Single Cell Transcriptome Analysis in Prostate Cancer

Chris Harris

Submitted to the University of Otago
in partial fulfilment of the
requirements for the degree of
Master of Science,
Genetics Major
2019

ABSTRACT

Background:

Single cell transcriptome studies have recently advanced to the analysis of millions of cells in a single pipeline. This type of analysis requires expensive, high maintenance platforms, precluding its use in most laboratories. Therefore, there is a need to find innovative ways of using more accessible low throughput methods for clinical application. Understanding the gene expression of an individual cell has clinical applications ranging from improved diagnosis to more precise treatment. We developed a novel single cell transcriptome analysis method and adapted it for ultra-low input mRNA analysis to potentially improve cancer diagnosis.

Methods:

Our single cell transcriptome analysis method allowed us to determine if the detection of single prostatic cells in various backgrounds could identify novel quantitative PCR biomarkers that could be used in urine diagnosis of prostate cancer. We applied this method to detect single cells from the LNCaP and PC3 prostate cancer cell lines in populations of prostatic (LNCaP) and non- prostatic (HeLa) cells. We used up to 29 cells from the HeLa cell line as a non-prostatic background for detection of a single LNCaP and PC3 cells. We also used up to approximately 199 LNCaP cells as a prostatic background to detect a single PC3 cell.

Results:

Our method detected 23 gene isoforms that significantly distinguish ($P < 0.05$) a single PC3 cell from a background population of approximately 199 LNCaP cells. 161 gene isoforms significantly distinguished ($P < 0.05$) a single PC3 cell from a background population of 29 HeLa cells. 64 gene isoforms significantly distinguished ($P < 0.05$) a single LNCaP cell from a background population of 29 HeLa cells. The MAGED1 gene had the highest log fold change in PC3 cells compared to HeLa cells (log2 fold change = 9.3, $P < 0.05$) and significantly distinguished ($P < 0.05$) a single PC3 cell from a background of 29 HeLa cells. qPCR analysis

confirmed the RNA-seq results for MAGED1 and PSA gene expression patterns in LNCaP and PC3 cells.

Conclusions:

Transcriptome wide single cell analysis identified several genes that indicate the presence of single PC3 and LNCaP cells in background populations of cells. This result supports a single cell sampling approach to highly sensitive diagnostics and prognostics for ultra-low input RNA samples. Our single cell analysis method compares positively with methods described in the literature with respect to genes detected per read.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ºC = Degrees Celsius

µg =Microgram(s)

µL = Microlitre(s)

µM = Micromole(s) per litre

aRNA = anti-sense ribose nucleic acid

ATCC = American Type Culture Collection

AD = Androgen dependent

AI = Androgen independent

bp = basepair(s)

BAM = Binary alignment map

CEL-seq = Cell expression by linear amplification and sequencing

CI = Confidence interval

CML = Chronic Myeloid Leukemia

Ct = Cycle threshold

CTCs = circulating tumour cells

cDNA = complementary DNA

DMEM-F12 = Dulbecco's modified Eagle's medium and F12 medium

DNA = Deoxy-ribose nucleic acid

DRE = Digital rectal examination

EDTA = Ethylenediaminetetraacetic acid

ERCC = External RNA Controls Consortium

FACS = Flow-assisted cell sorting

FBS = Foetal Bovine Serum

FU = Fluorescent units

G+C = Guanine-Cytosine

IVT = *In vitro* transcription

k = *a priori* number of assumed clusters for PAM analysis

RNA-seq --- RNA sequencing

kb = Kilobase(s)

L = Litre(s)

LNA = Locked nucleic acid

Log = Natural logarithmic value

Log2 = Logarithmic scale base 2

M = Mole(s) per litre

min = Minute(s)

mL = Millilitre(s)

MMLV = Moloney murine leukemia virus

MIPS = Michigan Prostate Score

mRNA = Messenger ribose nucleic acid

NCCs = Neuro-endocrine cancerous cells

NIST = National Institute of Standards and Technology

nt = nucleotides

ng = Nanogram(s)

nM = Nanomoles per litre

No. = number

Oligo-dT = poly-thymidine oligonucleotide

PC = Prostate cancer

PCA = Principal components analysis

PCR = Polymerase chain reaction

pg = Pictogram(s)

Poly-A = Poly-adenosine

PSA = Prostate specific antigen

PAM = Partitioning around the medoids

PBS = Phosphate-buffered saline

PCFNZ = Prostate Cancer Foundation of New Zealand

qPCR = quantitative PCR

RCD = Dr. Robert Day

RNA = Ribose nucleic acid

RPMI = Eagle's minimum essential medium

RT = reverse transcription

SAM = Sequence alignment map

sec = Seconds

STRT = Single cell tagged reverse transcription

T7 = T7 RNA polymerase

TRUS = Trans-rectal ultra sound

TS = Template switch

TSO = Template switch oligonucleotide

WHO = World Health Organisation

Chapter 1

1. Introduction

Understanding the relationship between genotype and phenotype is the principle aim of biological science. A key tool to assist this understanding is the study of transcriptomes. The 100 trillion cells in a typical human adult may share nearly identical genotypes, however transcriptome information reflects a unique subset of active genotype in each cell at a given time. Furthermore, due to the diverse array of cells that make up an organism, organ or tissue, bulk analysis of tissues omits biological differences between individual cells (Figure 1.1). These individual cell differences may have organism level consequences. Single cell transcriptome analysis is therefore an essential approach to functional studies of biological systems.

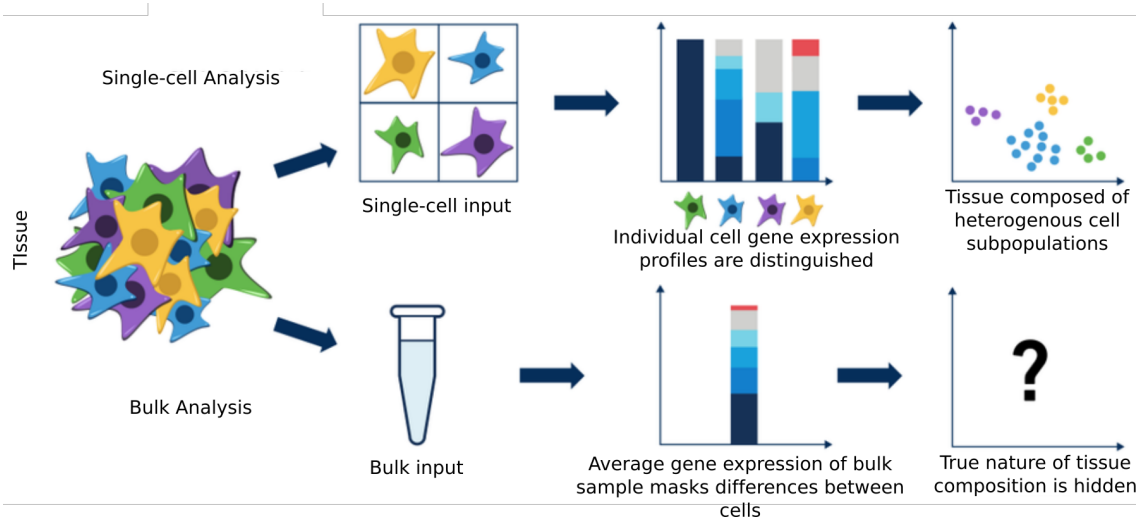

Figure 1.1 – Single cell vs. bulk cell analysis
Single-cell RNA-seq reveals cellular heterogeneity that is masked by bulk RNA-seq methods.

Single cell transcriptome analysis has advanced from the study of a handful of cells to high through put analysis of millions of cells in a single analytical pipeline (Svensson *et al*., 2018). This advancement in technology has taken place in the last 5 years, generating terabytes of data in a relatively short period of time (Figure 1.2). In 2016, the Human Cell Atlas Consortium emerged from these advances to discover, map and describe all human cell types. This represents the latest in a long line of consortia punctuating each major advancement in biology, beginning with the Human Genome Project. *Science* magazine

hailed high throughput single cell RNA-seq as the 2018 “Breakthrough of the year” for its ability to track early development.

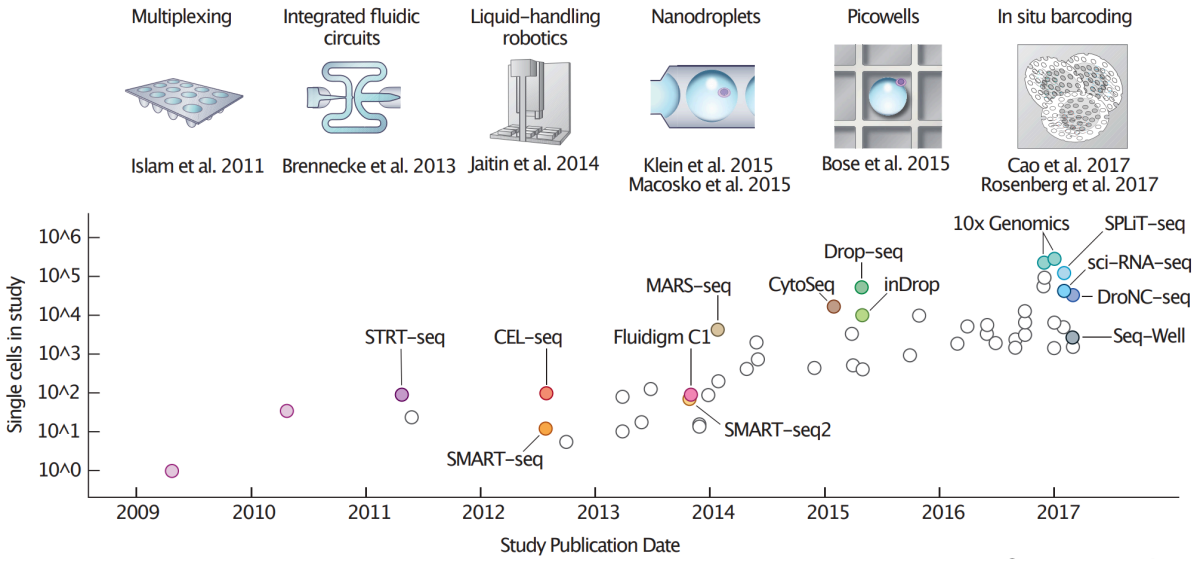

Figure 1.2 – Development of single cell transcriptome analysis

The first major jump in single cell processing volume was achieved through multiplexing by Islam *et al.* (2011), with STRT-seq. Fluidigm C1 was the earliest commercial microfluidic tool to isolate single cells and combined the SMART-seq reaction protocol to prepare multiplexed libraries for Next Generation Sequencing (NGS) analysis. Subsequent microfluidic tools, such as inDrop and Drop-seq, used nano-litre droplets to isolate single cell reactions. The 10X Genomics platforms are the latest iteration of oil droplet microfluidics (Svensson *et al.*, 2018).

The effort to identify rare cell types has emphasised the importance of maximising the number of single cell transcriptomes analysed. Identifying rare cell types has significant implications for immunology and oncology, where rare cell types can have significant consequences for patient outcomes. Tumours and the immune system are complex ecosystems due to their high connectivity, heterogeneity and plasticity. External cell markers have proven inadequate in delineating cell types in these systems (Hume *et al.*, 2008; Lawson *et al.*, 2018). Therefore, high-through put single cell transcriptome analysis is the only effective tool for comprehensive cellular segregation of complex tissues without the need for predefined markers (Giladi and Amit, 2018).

In the drive to improve cell number through put, single cell transcriptome analysis has transitioned away from laboriously depositing single cells in plated wells, to using microfluidic cell technology to separate cells and complete the amplification reaction in a dynamic process. Currently, the most common single cell transcriptome library preparation

method is high through-put microfluidic technology (Kulkarni *et al*., 2019). The three most widely used platforms are 10X Genomics Chromium, Drop-Seq and inDrop. These approaches use microfluidics to tag individual droplets containing an individual cell with a unique barcode. Within each cell, unique molecular identifiers also tag each mRNA transcript, allowing for the removal of technical biases.

Consequently, low-throughput methods have been over-shadowed by high-throughput methods. The remaining advantages of low-throughput well based methods is that they use ubiquitous skills, equipment and limited infrastructure. Any laboratory with RNA-seq analysis capabilities may use these methods without a significant capital investment. The essential components include 96 well plates, reagents for mRNA amplification and a method for isolating single cells and RNA-seq technology. Recasting low-throughput single cell technologies for studying ultra-low input mRNA could lead to their re-establishment, particularly in the context of poorly resourced research environments.

In 2015, Day *et al.* (2018) developed a well-based single cell transcriptome method that is scalable to potentially hundreds of cells in a single analytical pipeline. We applied a part of the Day *et al.* method herein to analyse ultra-low input RNA samples and single cell transcriptomes. The method is comparable to SMART-seq2 because it is well-based and involves a PCR amplification to create a cDNA library amenable to RNA-seq analysis (Picelli *et al.,* 2014)). In terms of raw numbers of cells analysed, this technology cannot compete with current high-throughput droplet based microfluidic technologies.

Day *et al.*'s method has the advantage over SMART-seq2 with regards to early multiplexing of single cell samples, where each well in a 96-well plate is assigned a 5' barcode before library amplification. Although this method is relatively low throughput, we demonstrate that it remains a useful tool for cheaply interrogating single cells and micro-populations of cells to answer contemporary biological questions. One of the novel questions this thesis aims to address is the ability of a single cell RNA-seq method to analyse a population of cells and detect a single cell transcriptome of interest in that population. This application of a single cell technology to micro-population analysis could encourage re-adaption of low

throughput single cell technologies to novel contexts outside of analysing individual single cells.

Given the potential of our method to analyse multiplexed micro-populations of cells, this allows our method to process cells in the order of thousands in one pipeline. This may be of clinical relevance to urine samples, where a patient may shed upwards of 6000 intact cells within a 24-hour period (Lang *et al.*, 2013). For example, very early stage prostate cancers may shed very small numbers of cancer cells in urine, particularly after intensive prostate massage (Mengual *et al*., 2016). Distributing urinated cells across 96 well plates could allow a sensitive single cell method to detect a single shed cancerous or pre-cancerous cell amongst small sub-populations of cells in each well.

There are presently diagnostic tools available to detect cell free mRNA transcript disease biomarkers in urine. However, combining a sensitive single cell transcriptome method that can detect low abundance transcripts with well plate based segregating of micro-populations of urinated shed cells may offer advantages. Such a method could detect significantly more *de novo* mRNA species with diagnostic significance than using a pre-defined suite of qPCR targets. Ultra-low input mRNA that can be analysed directly after cell lysis means that the risk of RNA degradation is lowered. mRNA integrity can be maintained for hours when it is protected by cellular membranes (Meng *et al*., 2012; Fujita *et al.,* 2009; Quek *et al.,* 2012).

We developed an experimental design to test the hypothesis that a single cell method can be used to interrogate small populations of cells to identify a single cell transcriptome of interest. We used the PC3 and LNCaP prostate cancer cell lines as a model of uro-genital disease. We compared our single cell data with analogous data from SMART-seq2, another well based method used as standard practice in many laboratories (Baran-Gale *et al*., 2018) These cell lines were also used for model single cells in the paper describing SMART-seq2 (Picelli *et al*., 2014).

1.1 Overview

The first chapter of this thesis is a survey of the latest developments in single cell transcriptome analysis methodology, including computational analysis and current challenges in the field. We briefly outline the thesis objectives, our single cell transcriptome analysis methodology and experimental design. We discuss the thesis method in the modern context. The final section describes the modern computational analysis methods and our own computational approach for analysis of our single cell transcriptome data.

The second chapter is a detailed description of the thesis single cell transcriptome methodology and experimental design. The computational analysis and normalisation method of the experimental data is also outlined. Finally, qPCR experiments ancillary to our initial findings are described.

The third chapter is a presentation of the RNA-seq data generated by our single cell transcriptome platform and a comparison with analogous data from Picelli *et al.* (2008) generated by a different methodology. The data described originated from single PC3 and LNCaP cells. We present qPCR validation of the RNA-seq data generated by our method. We also describe data from the application of our platform to detecting the transcriptome of single cells in a dissimilar cell population.

The final chapter is a discussion of our data in the current context of single cell technology literature. We critically evaluate the advantages and disadvantages of our method and experimental design. We also review the challenges in computational analysis. Finally, future directions of the single cell transcriptome field are discussed.

1.2 Single cell transcriptome sequencing methods

Although single cell RNA-seq technologies have advanced profoundly in terms of through put, they still retain the basic features of all single cell analysis methods. There are three broad steps in common between microfluidic and well based approaches:

i. Isolating the cells of interest from a background of uninteresting cells;
ii. Reverse transcription of mRNA to double stranded cDNA; and

iii. Amplification of the cDNA to levels amenable to analysis, whether by qPCR (end-point analysis) or further whole transcriptome sequencing.

Microfluidic technologies such as the single cell controller platform from 10X Genomics encapsulate single cells into nanolitre droplets containing DNA barcoded reads for reverse transcription. After reverse transcription, the cDNA for each single cell is uniquely DNA barcoded and can multiplexed in a single RNA-seq run. These high through-put technologies can channel millions of individual cells through a microfluidic chip with compartments that isolate cells for the reverse transcription and amplification reaction.

Well based technologies such as SMART-seq2 can take advantage of micro-pipetting or fluorescence-activated cell sorting (FACS) to isolate single cells from tissues. Micro-pipetting is a non-biased and straightforward for cell culture after trypsinisation, which was the method for isolating single cell in this thesis. FACS can introduce biological bias from transcriptomes being linked to the cell surface protein markers, cell size and cell cycle status. Single cells can then be deposited into 96 or 384 well plates containing a cell lysis buffer and excess ribonuclease inhibitor to protect the RNA. This type of cell isolation requires the cells to be processed quickly with all downstream work carried out on ice.

Both the microfluidic Genome 10X approach and the well plate based methods enrich for coding mRNA by using an oligot(dT) containing primers for RT and PCR. This is possible because oligo(dT) anchors to the poly adenine (poly A) tail of mature mRNA, but may also anchor to internal poly A regions of mRNA transcripts.

1.2.1 Droplet microfluidic technologies

Single cell microfluidics is primarily a method of isolating large numbers of individual cells for transcriptome analysis. Fundamentally, it involves isolating single cells inside oil droplets through microfluidic channels. In the case of Genome 10X platforms, this allows a compartmentalised reaction to attach a unique gel bead barcode for each single cell transcriptome (Figure 1.3). This process forms gel beads in emulsions (GEMs), where barcoded individual transcriptomes can be pooled for more efficient downstream reactions.

The Genome 10X single cell platforms have been the most prolifically utilised microfluidic tools in recent years. Over 1200 instruments have been sold to 93 out of 100 of the top research institutions by publications, generating over 500 publications since the single cell series was released in mid 2015. The defining advantage of the 10X system is its ability to process many cells in a short space of time. It can capture 100-10,000+ cells per output channel and up to 80,000 cells per run in < 7 minutes. Although popular, wilful intellectual property violations have resulted in court injunctions preventing further sale of this system. This has resulted in Genome 10X releasing a new platform in late 2019 with a similar cell processing capability.

The purpose of the Genome 10X single cell system is to create an emulsion of oil droplets, each containing a single cell (Figure 1.3). mRNA in each droplet is reverse transcribed to cDNA containing a unique molecular barcode attached to a gel bead. The barcode allows for downstream pooling of samples and bulk amplification by PCR followed by bulk NGS. The emulsion is created by a chip containing eight units of wells and channels. Each unit comprises three wells feeding a network of channels leading to an outlet well for the emulsion. Each set of wells includes: (i) a sample well containing disassociated cells and reagents for reverse transcription; (ii) a well containing the droplet forming oil; and (iii) a well containing the barcoded beads.

The contents of the input wells combine in the channel network to release a functional GEM containing a single cell, a single gel bead, and RT reagents. Within each GEM reaction vesicle, a single cell is lysed, the gel bead is dissolved to free the identically barcoded RT oligonucleotides into solution. Reverse transcription of mRNA subsequently occurs, resulting in all cDNAs from a single cell attached to the same barcode. This allows sequencing reads to be indexed to their original single cells.

While the 10X single cell system has significant advantages in terms of high volume single cell analysis, it has some drawbacks (Luo *et al.,* 2019). Single cells and gel beads are randomly encapsulated in each droplet and the number droplets containing cells follows a non-uniform poisson distribution. This means very few droplets will contain both a barcoded bead and a single cell. Cell capture rates vary from 50-65%, making it more

difficult to capture rare cell types. In addition, some cell types adhere very tightly together, resulting in some GEMs containing multiple cells. complicating computational single cell analysis. These ‘mulitplets’ are likely to add significant noise to single cell computational analysis. Cells derived from trypsinised cell culture or liquid biopsies are more likely to produce a truly single cell suspension than cells derived from solid tissue (O’Flanagan *et al*., 2019).

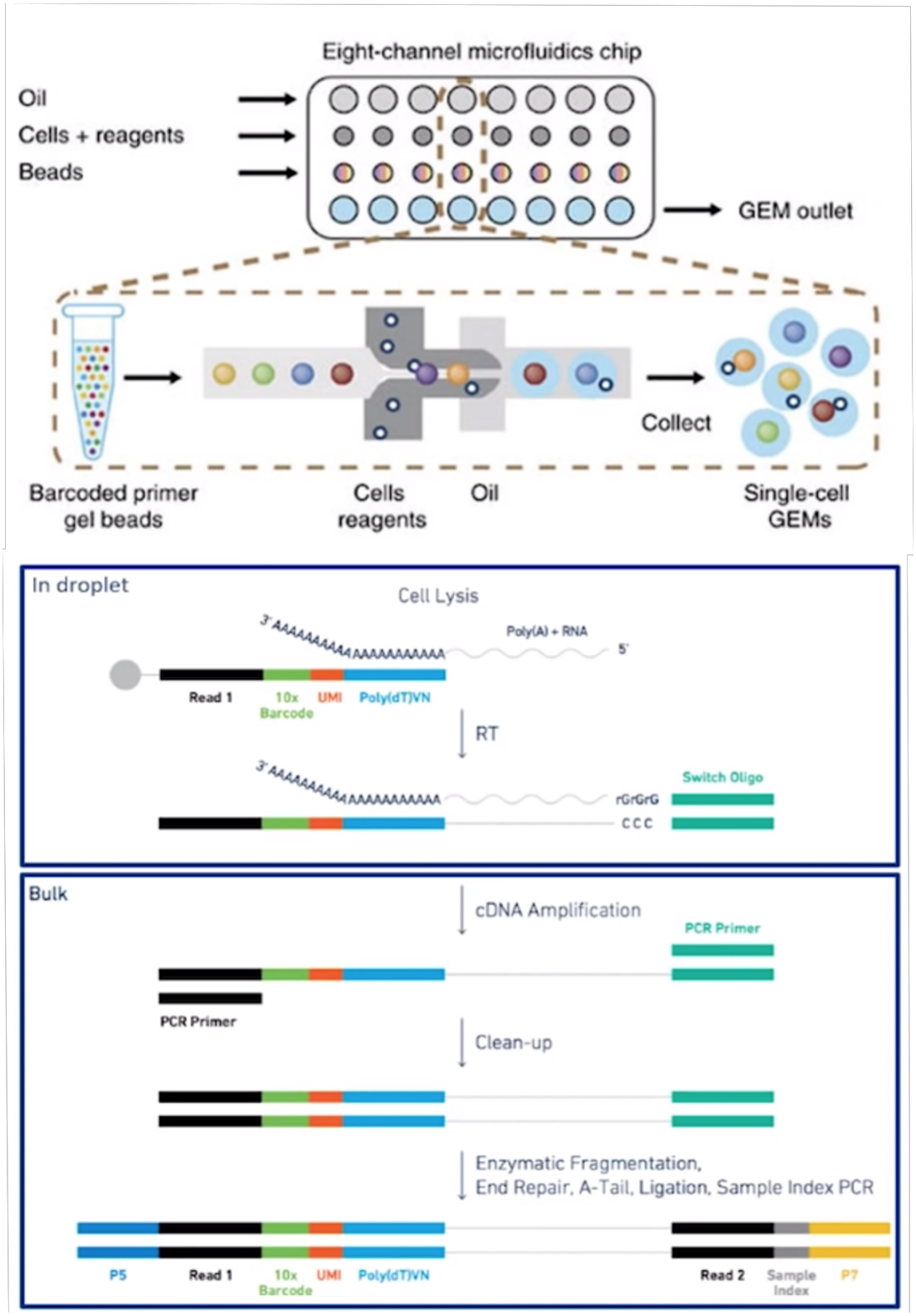

Figure 1.3 – Genome 10X single cell system

The eight channel microfluidics chip permits the isolation of a large number of single cells by oil droplets formed around a cell, reagents and a barcoded primer gel bead. The RT reaction takes place inside each droplet to incorporate the 10X barcode for downstream aggregation of cells for an efficient single bulk cDNA amplification reaction. Genome 10X is based on a 3' tag sequencing method.

1.2.2 SMART-seq2

SMART-seq2 is the commercial standard plate based method for single cell analysis currently offered by Illumina (Picelli *et al.,* 2014). Both SMART-seq2 and the Genome 10X method exploit the SMART reaction (Switching Mechanism at the end of the 5'-end of the RNA transcript). This template switch mechanism enables the incorporation of forward and reverse PCR primers following reverse transcription of the mRNA.

Developed by Matz *et al.* (1999), the template switch procedure utilises the ability of the Moloney Murine Leukemia Virus (MMLV) reverse transcriptase to add a few non-templated cytosines to the 3' end of the newly synthesized cDNA strand (Figure 1.4). This usually occurs when the reverse transcriptase reaches the capped 5' end of the mRNA. These extra cytosines work as a docking site for the three riboguanosines attached to a 'Template Switching Oligonucleotide' (TSO). The reverse transcriptase is then able to 'switch template' (from mRNA to the DNA of the TSO) and synthesize a complementary DNA strand using the TSO as template (Picelli, 2017).

SMART-seq2 is the second generation of this method, developed in in 2013 by Picelli *et al*. The TSO in the Smart-seq2 method replaces the terminal riboguanosine with a locked nucleic acid (LNA)-modified de-oxyguanosine (Figure 1.4). Locked nucleotides are characterized by an internal bond between the O2' and the C4' of the furanose ring, linked by a methylene group (Picelli, 2017). The modification introduces a conformational restriction in the molecule, which still retains the physical properties of the original nucleic acid. Two properties of LNAs are advantageous for template switching: the enhanced thermal stability of the LNA monomers and their ability to anneal strongly to the untemplated 3' extension of the cDNA.

SMART-seq2 also improved the lower read coverage towards the 5' end of transcripts that was characteristic of full length coverage methods up to that point. The addition of $MgCl_2$ and betaine to the template switch reaction assisted in processivity of the DNA polymerase (Picelli *et al.,* 2013) The final and perhaps most important advantage of SMART-seq2 is that

it entirely relies on off-the-shelf reagents, reducing the capital investment required compared to microfluidic methods.

SMART-seq2 has some important limitations depending on the biological hypothesis to be tested. Samples can only be pooled immediately prior to sequencing, making the method more labour intensive than barcode tag-based methods such as STRT-seq (Islam *et al*., 2014). Despite being a well based approach with lower cell through-put, SMART-seq2 has important advantages compared to Genome 10X microfluidic technologies. Smart-seq2 is able to detect more genes per cell than Genome 10X, making it a more appropriate tool for experiments requiring high sensitivity (Wang *et al.,* 2019).

Well based methods such as SMART-seq2 have attributes that continue to make them relevant to current biological questions related to ultra-low RNA inputs. A modified SMART-seq2 method was recently used to investigate endogenous retroviral differential gene expression between low numbers of immature and mature mouse oocytes (Treger *et al*., 2019). This approach enabled sensitive detection of low-abundance retro-element transcripts in growing oocytes and to detect differential expression between sub-strains. The full-length read coverage of SMART-seq2 also enabled improved mapping accuracy of repeat elements compared to a 3' tag sequencing method used by Genome 10X (Treangen and Salzberg, 2011). The low cost of SMART-seq2 was also noted in the study.

Unlike Genome 10X, SMART-seq2 analyses the full length of the mRNA transcript by using a tagmentation reaction. This precludes the use of unique molecular identifiers (UMIs) for molecular counting. Molecular counting enables the calculation of the absolute numbers of each original mRNA transcript. Islam *et al.* (2014) determined the absolute quantity of original mRNA transcripts by tagging each mRNA molecule with a UMI (a random 5bp sequence) before amplification. The quantity of original molecules for each mRNA species should therefore correspond to the quantity of unique molecular identifiers present in the amplified population for each mRNA species.

The tagmentation reaction that introduces sequencing primers to transcripts for next generation sequencing converts full cDNA strands into multiple fragments (Figure 1.4).

Therefore, the UMI barcode would be lost because it can only label the terminal portion of the full-length transcript. Molecular counting of fragmented transcripts is still possible with ERCC spike ins.

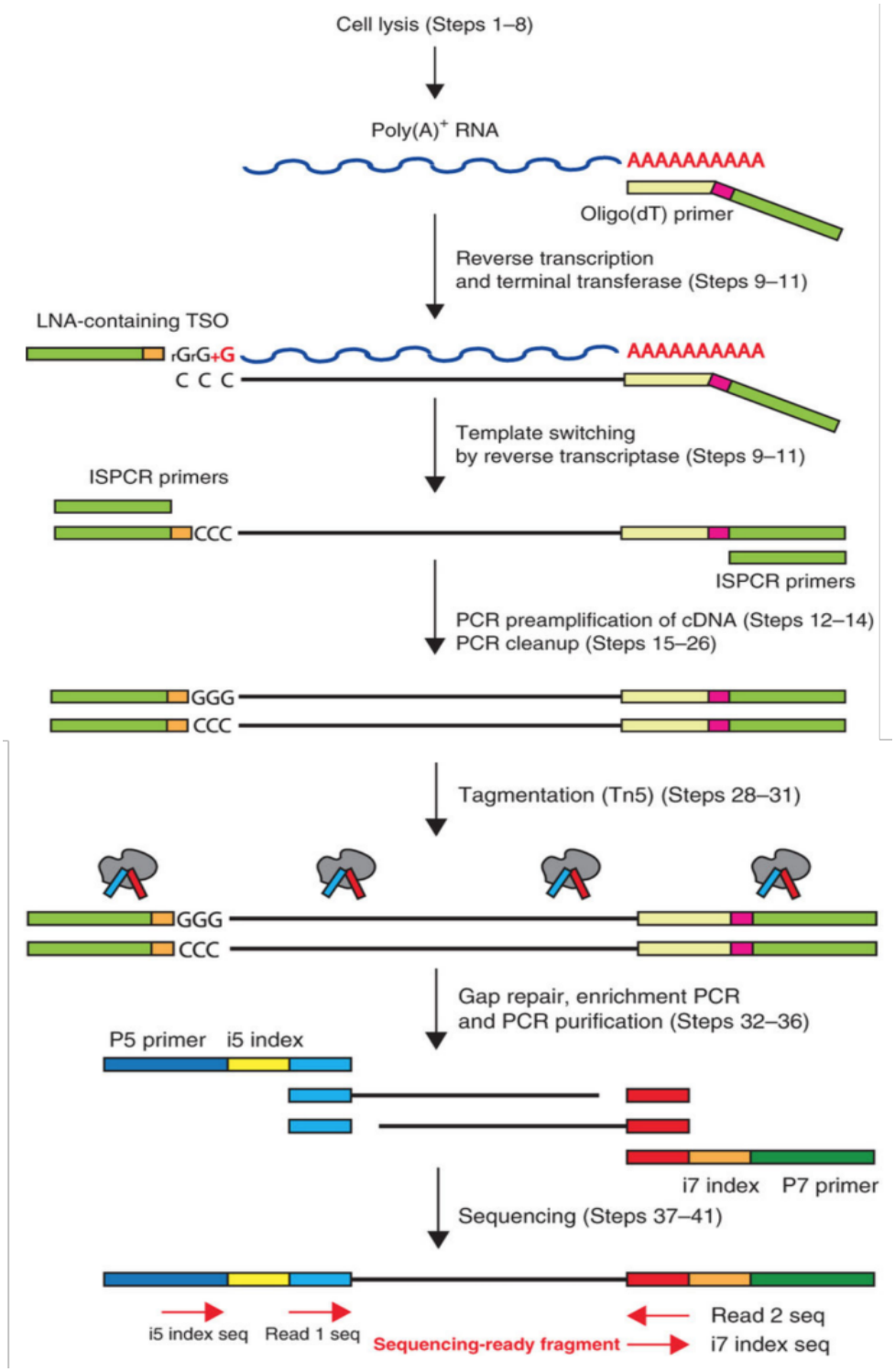

Figure 1.4 – Smart-seq2 protocol

Outline of the SMART-seq2 protocol and the corresponding procedure steps (Picelli *et al.*, 2014). Poly A mRNA is reverse transcribed from an oligo(dT) primer. The MMLV derived reverse transcriptase adds 2-5 cytosines to the newly synthesised cDNA (3 in figure). The LNA-containing TSO carries 2 riboguanosines (rG) and a LNA modified guanosine (+G). Both the oligo(dT) and LNA-containing TSO contain palindromic PCR primers (ISPCR primers) to enable PCR amplification of the cDNA. The amplified cDNA is then prepared for sequencing by tagmentation via Tn5 transposase. The cDNA is fragmented and another set of PCR primers are attached to each fragment for further PCR, followed by addition of sequencing primers and a barcode (i5 and i7 indices).

1.3 Aims

The overall aim of this thesis was to troubleshoot and test a novel well based single cell transcriptome method, and subsequently apply it to analysis of a bulk cell micro-population context to determine if the method could delineate a single cell in a bulk sample. In testing the single cell method, we used the method to analyse the single cell transcriptomes of two prostate cancer cell lines and compare the data generated to SMART-seq2 data generated from the same cell lines. This comparison enabled us to have a point of reference for evaluating both methods.

After establishing a method capable of analysing whole mRNA from single cells and ultra-low inputs, we created an experimental design to detect a single cell in a bulk sample. The purpose of this aim was to see if the method could be applied to detect a large number of mRNA species from a single cell of interest in micro-population samples, with a view towards future diagnostic applications in a cancer fluid biopsy context. To this end, we created various combinations of two prostate cancer cell lines and the HeLa cell line with a single cell deposited in a background.

The biological question which we want to address is the sole determinant of the appropriateness of the technique we wish to use, irrespective of advancements in through-put. We suggest that a well plate based technique is more appropriate than a microfluidic technique high through put technique because it remains capable of answering the specific biological question we wish to address and is much more accessible in terms of cost.

1.4 Thesis method and design

This thesis utilised part of an innovative well plate based method described by Day *et al.* (2018) that can be used for single cell transcriptome and ultra-low input RNA analysis. The method was coupled with an experimental design to analyse ultra-low input mRNA derived from a micro-population of cells to detect a single cell transcriptome amongst a population. The method adapts features from other plate based single cell analysis methods. Day *et al.* (2018) incorporate elements such as template switching, PCR amplification and tagmentation from SMART-seq2 and barcoding of individual cells by the oligo(dT) from

STRT-seq (Picelli *et al*., 2014; Islam *et al*., 2011). This early barcoding enables downstream pooling of samples post RT to reduce the costs and labour of sample amplification.

The Day *et al.* method also incorporates an *in vitro* transcription (IVT) step. IVT was first described by Eberwine *et al.* (1992) as a method to linearly amplify cDNA derived from single cells. We utilise IVT post PCR to focus our analysis on the 3' end of mRNA transcripts, eliminating the need for gene length correction for data normalisation. Linear amplification methods for single cells have been largely based on the T7 RNA polymerase (T7), derived from bacteriophage. The T7 catalyses the formation of first strand anti-sense RNA (aRNA) from template cDNA (known as *in vitro* transcription or IVT). T7 can amplify RNA up to 1000 fold in one round of amplification (Wilhelm *et al.,* 2006). It has the advantage of being extremely specific to a T7 promoter sequence. Limitations include the low processivity of T7 and a requirement for double stranded promoter sequence to initiate T7.

The inherent limitation of single cell IVT is that at least 400pg of starting RNA is required. Hashimshony *et al.* (2012) remove this problem by pooling samples that have been barcoded with a unique sequence in the Oligo-dT after the reverse transcription while also providing a multiplex capability similar to the PCR STRT method. We similarly overcome this hurdle by pooling samples post PCR into a single aliquot for IVT amplification. Hashimshony *et al.* (2012) compared the data gathered from their method with the data from Islam's PCR method, and found their linear method was more reproducible, amplified more genes than STRT, and was better able to distinguish gene expression levels between different cell types.

1.4.1 Experimental design

We designed three experiments in order to address the aim of detecting mRNA transcripts from a single cell in a micro-population. We deposited either a single PC3 or LNCaP cell in background populations of either approximately 29 HeLa cells or 199 LNCaP cells. As illustrated in Chapter 2, we compared the samples with a deposited single cell with sample composed of entirely of the background population with a similar number of cells to carry out differential gene expression analysis with an empirical Bayes method.

1.5 Prostate cancer cell lines

Prostate cancer is the most common cancer in men and has significant challenges in diagnostic and prognostic accuracy (Catalona *et al.*, 2017). The development of more specific biomarkers for high-risk prostate cancer is necessary, because the prostate-specific antigen (PSA) blood test lacks specificity for the detection of prostate cancer and can lead to unnecessary prostate biopsies (Siegel *et al.,* 2013; Ferlay *et al.,* 2010).

There is scope for ameliorating diagnostic specificity and sensitivity through improving analysis of cellular components of prostate cancer sloughed into urine. A great improvement in high-throughput gene expression techniques has yielded several promising molecular biomarkers for prostate cancer detection (Mengual *et al*., 2016). This has potential for improving multiplexed qPCR assays for urinary RNA. Testing our experimental design on different prostate cancer cell lines is a logical initial step with a view towards analysing micro-population samples in patient urine.

The cell types we used were determined by both the potential future clinical applications, availability, and what cell lines had been studied in the literature. We applied our method and experimental design to the PC3 and LNCaP prostate cancer cell lines. These two cell lines broadly model the two major types of prostate cancer in terms of aggressiveness and cell differentiation. They have been used in previous single cell transcriptome studies (Picelli *et al*., 2014).

Prostate cancer can be roughly divided into two subtypes: androgen dependent (AD) and androgen independent (AI). Early stage prostate cancers tend to be AD. AD prostate cancers require androgen receptor signaling activated by testosterone derivatives for growth (Karantanos *et al.*, 2013).

Chemical castration reduces the amount of testosterone in the blood, resulting in shrinkage of AD prostate tumours. This effect is usually temporary, with AI prostate cancer emerging after a mean time of 2-3 years. The reasons for this are largely unknown, but recent investigations support the theory that androgen removal provides a selective advantage to

AI cells, which eventually repopulate the tumour (Taplin *et al.,* 1999; Craft *et al*., 1999; Culig *et al*., 1998; Li *et al*., 2001).

AI prostate cancer can be caused by mutations to the androgen receptor that increase sensitivity to the few testosterone derivatives remaining or mutations that broaden the specificity of AR, so that it can be activated by non-androgenic molecules. PSMA (FOLH1) gene expression has been implicated as providing an alternative mechanism to androgen driven growth by increasing folic acid levels, an important molecule for rapidly dividing cells (Yao *et al.,* 2010).

The LNCaP cell line is an AD cell line that requires the androgen present in foetal bovine serum cell culture medium for growth. LNCaP was established from a metastatic lesion of prostate adenocarcinoma from a 50-year-old Caucasian male (Chu *et al.,* 1983). LNCaP tends to express androgen regulated genes such as PCA3 and PSA (Ferreira *et al*., 2012).

PC3 is an androgen insensitive cell line and is therefore lacking in PSA and PCA3 expression. PC3 was established from a bone metastasis of Grade IV metastatic prostate adenocarcinoma from a 62 year old Caucasian male (Kaighn *et al.*, 1979). PC3 is not considered clinically relevant because it forms osteolytic lesions *in vivo,* whereas clinical prostate cancer bone metastasis is osteoblastic. LNCaP tends to express PCA3 and PSA, whereas PC3 does not.

1.5 Computational analysis

A significant part of this thesis was devoted to computational analysis of the data generated by the single cell and micro-population experiments. There are significant computational challenges in single cell analysis related to data normalisation and cell type identification. Several computational pipelines exist for single cell analysis, but gold standard tools have yet to be developed (Hwang *et al*., 2018).

One of the major challenges in single cell RNA-seq analysis is data normalisation. The purpose of data normalisation is to remove the effects of systematic technical variability

and unwanted biological variability in gene expression, so that biological differences of interest can be observed. Several normalisation techniques for single cell RNA-seq have been proposed, including scaling methods (Lun *et al*.,2016), regression based methods for known nuisance factors (Buettner *et al*., 2015) and methods that rely on spike-in sequences from the External RNA Controls Consortium (ERCC).

We describe a data normalisation protocol in Chapter 2 and discuss the relevant literature with regards to our method in Chapter 4. We aimed to remove variation between single cells attributable to technical biases and identify any significant bias due to the cell growth cycle. In Chapter 2 we also illustrate how our experimental design uses computational analysis of differential gene expression between a study population and a homogenous control population to identify several mRNA transcripts from a 'rare' single cell in a population background. We utilised an empirical Bayes approach for differential gene expression analysis.

Apart from data normalisation, we considered issues related to using clustering algorithms to identify single cell types. To this end, we primarily reduced the thousands of dimensions involved in a transcriptome to the first two principal components. This was used in conjunction with partitioning around the medoids (PAM) clustering to distinguish single cell types. Principal components analysis (PCA) is a common linear method for simplifying single cell datasets. In contrast to non-linear methods such as t-SNE, linear methods tend to preserve global data structure and are better at separating clusters (Nguyen *et al*., 2019).

In Chapter 4 we discuss how the attributes of an empirical Bayes framework are useful for single cell analysis in the context of mitigating outlier effects, technical bias and possibly biological bias stemming from the cell cycle effect. An empirical Bayesian approach to single cell data normalisation was recently described by Tang *et al*. (2019). Empirical Bayes makes use of global scaling factors that corrects technical variations including differences in capture efficiency (Catalina *et al*., 2017). Also, global scaling factors correct for biological variations due to difference in transcript content and cell size.

Chapter 2

2. Materials and Methods

2.1 Materials

2.1.1 Reagents

0.05% trypsin solution (prepared in lab, Appendix 5.8.2)

Ambion™ MessageAmp™ II aRNA amplification kit (Thermo Fisher Scientific, USA)

AMPure® XP magnetic beads (Thermo Fisher Scientific, USA)

Betaine, 5 M (Sigma-Aldrich, USA)

Dimethyl sulfoxide (DMSO – Sigma-Aldrich, USA)

dNTP Mix, 10mM (Invitrogen, USA)

Dulbecco's phosphate-buffered saline (DPBS – Sigma-Aldrich, USA)

Dulbecco's modified Eagle's medium and F12 medium (DMEM-F12 – Invitrogen, USA)

DTT, 100mM (Sigma-Aldrich, USA)

Eagle's minimum essential medium (RPMI-1640 - Invitrogen, USA)

ERCC Spike-in control mix, 1/1000 (Invitrogen, USA)

Ethylenediaminetetraacetic acid (EDTA - Invitrogen, USA)

Foetal bovine serum (FBS – Invitrogen, USA)

Isopropyl alcohol (Scharlau, Spain)

KAPA HiFi™ HotStart ReadyMix, 2× (Kapa Biosystems Inc., USA)

KAPA SYBR® FAST qPCR Master Mix Universal (Kapa Biosystems Inc., USA)

$MgCl_2$, 250 M (prepared in lab)

NEBNext® Fragmentation Buffer (New England Biolabs, USA)

Nextera DNA Library Preparation Kit (Illumina, USA)

Oligonucleotide dT30VN (RCD oligo-dT - Integrated DNA Technologies, USA)

PCR and qPCR primers (Integrated DNA Technologies, USA)

Phosphate buffered saline (PBS) solution (prepared in lab, Appendix 5.8.3)

PrimeScript™ Reverse Transcriptase, 200 U/µL (Takara Bio Inc., Japan)

PrimeScript™ Buffer, 5X (Takara Bio Inc., Japan)

RNase inhibitor (Qiagen, USA)

Template switch oligonucleotide (TSO) containing Biotin, 10µM (Integrated DNA Technologies, USA)

TURBO™ DNase (Thermo Fisher Scientific, USA)

Triton X-100 detergent, 10% (Sigma-Aldrich, USA)

Trypan blue (Bio-rad, USA)

UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen, USA)

2.1.2 Equipment

0.6mL microtubes (Axygen, USA)

1 mL cryotube (Nalgene Labware, USA)

1.5 mL microtubes (Axygen, USA)

10 mL Serological pipettes (Greiner Bio-One, Germany)

15mL Falcon tubes (BD Biosciences, USA)

25mL cell culture flasks (Greiner Bio-One, Germany)

50mL Falcon tubes (BD Biosciences, USA)

75mL cell culture flasks (Greiner Bio-One, Germany)

Agilent 2100 Bioanalyser High Sensitivity DNA Assay (Agilent Technologies, Germany)

Applied Biosystems GeneAmp^{TM} PCR system 2700 cycler (Thermo Fisher Scientific, USA)

Applied Biosystems 7900HT Fast Reat-Time PCR system (Thermo Fisher Scientific, USA)

Centra 3C centrifuge (International Equipment Company, USA)

CO_2 cell culture incubator (Binder, Germany)

Dual chamber cell counting slides (Bio-Rad, USA)

Express SC250EXP SpeedVac™ Concentrator System (Thermo Fisher Scientific, USA)

Milli-Q Ultrapure Water Purification System (Millipore, USA)

MiSeq System (Illumina, USA)

Mr. Frosty 5100 Cryo 1ºC Freezing Container (Thermo Fisher Scientific, USA)

Nanodrop ND-1000 Spectrophotometer (Nanodrop Technologies, USA)

Qubit High Sensitivity DNA assay (Thermo Fisher Scientific, USA)

RNeasy® MinElute® Cleanup spin columns (Qiagen, USA)

TC10 Automated Cell Counter (Bio-Rad, USA)

Tissue culture hood (EMAIL, Australia

Water bath (Semco, USA)

2.2 Single cell whole transcriptome analysis methodology overview

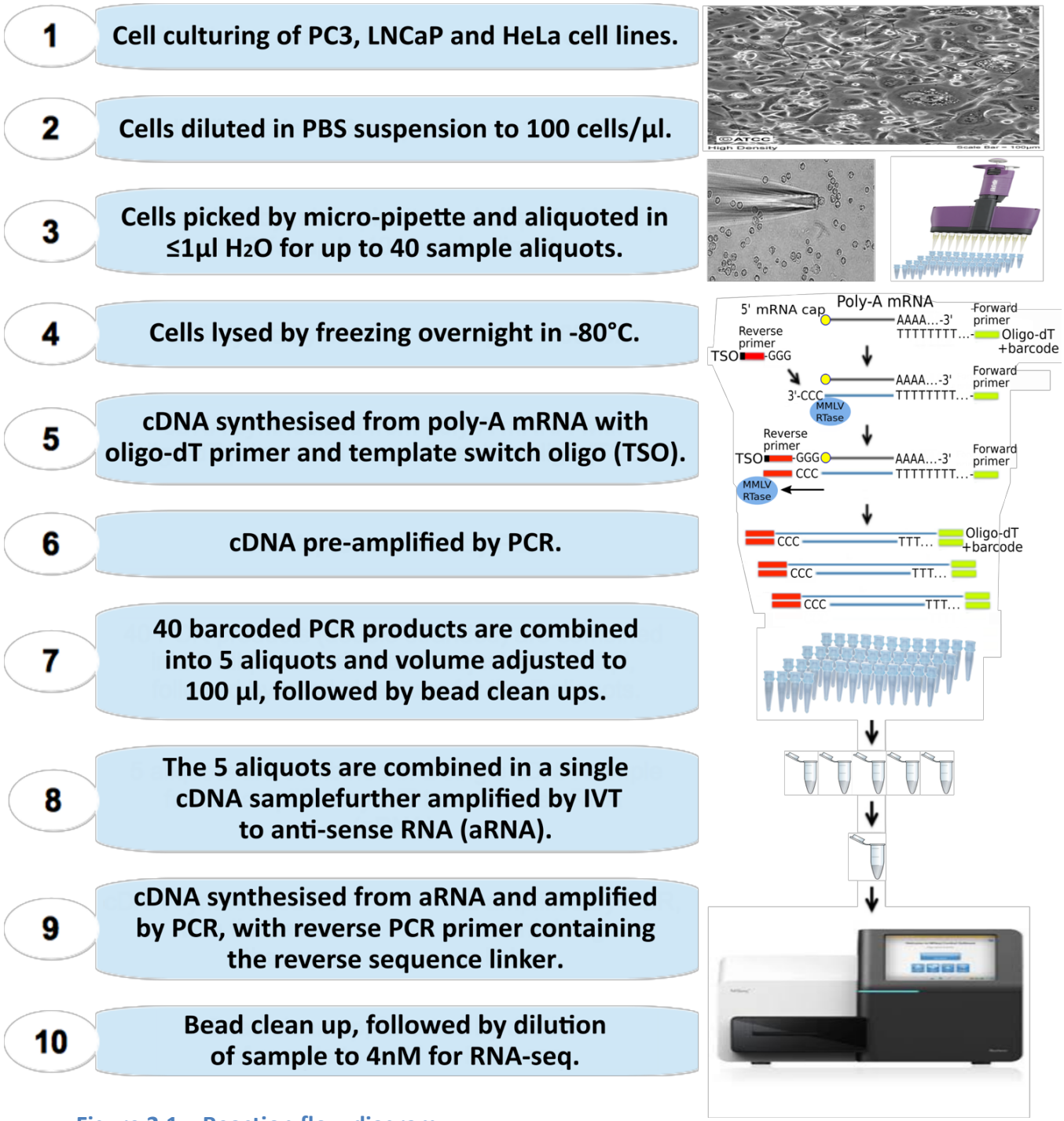

Figure 2.1 – Reaction flow diagram

Diagram depicting the basic steps of the method used in this thesis to produce single cell transcriptome data.

2.3 Experimental design

We applied our amplification method to three major whole transcriptome analysis experiments:

i. Detecting single PC3 cells in a background population consisting of 29 HeLa cells;
ii. Detecting single LNCaP cells in a background population consisting of 29 HeLa cells; and
iii. Detecting single PC3 cells in a background population consisting of 199 LNCaP cells.

The basic method of detecting a prostatic transcriptome signature in background populations consisted of first determining the up-regulated genes in a pure prostatic population (e.g. 30 PC3 cells) compared to a pure background population (e.g. 30 HeLa cells). We then sought to determine if these up-regulated genes (e.g. a PC3 or prostatic signature) were also up-regulated in the background populations spiked with a single prostatic cell (e.g. 1 PC3 cell in 29 HeLa cells) compared with pure background populations (e.g. 30 HeLa cells).

We surmised that the single prostate derived cell was detected in background if the prostatic signature was up-regulated in both comparisons (Figure 2.2). The genes that were up-regulated in both comparisons were considered for prostatic cell marker development in a quantitative PCR assay.

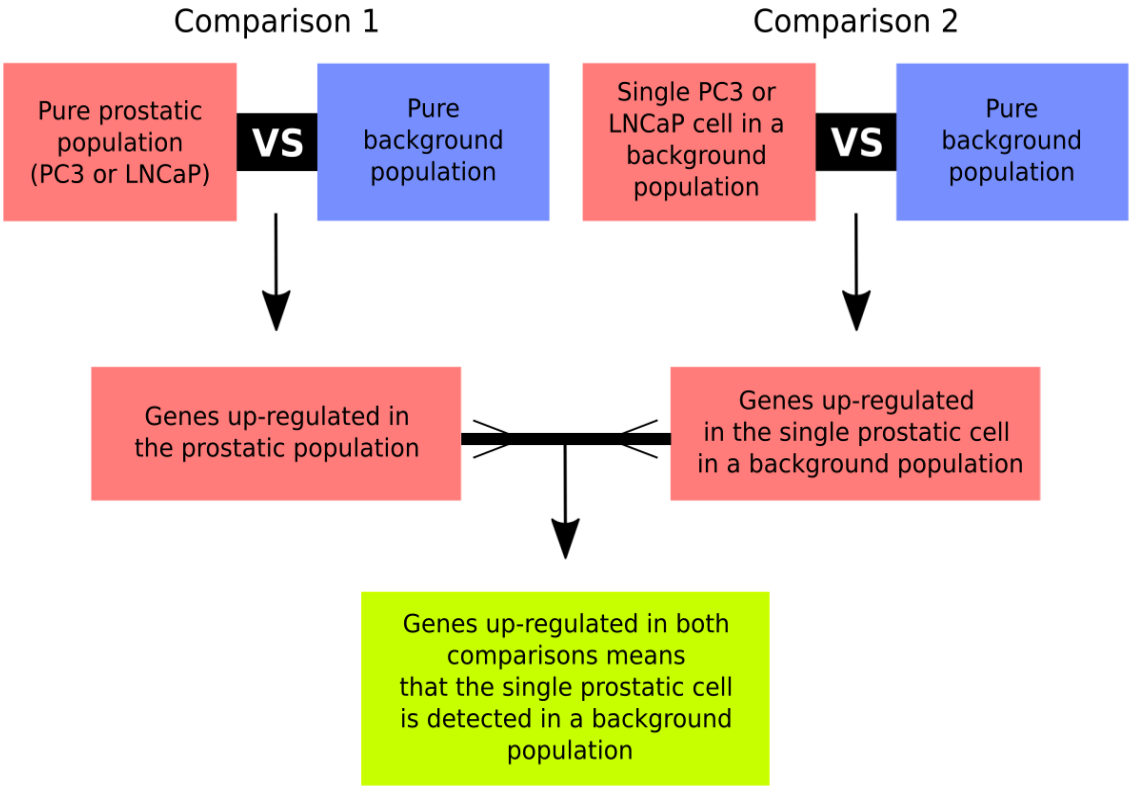

Figure 2.2 – Method for detecting single cells

Diagram illustrating the basic experimental design for detecting genes in a cell mixed with an unrelated background population

2.4 Cell lines

Human prostate cancer cell lines (LNCaP and PC3) and a cervical cancer cell line (HeLa) were cultured to study a single prostate cancer cell separately as well as in a background population of HeLa cells.

The LNCaP, PC3 and HeLa cell lines were purchased from the American Type Culture Collection (ATCC).

2.5 Complete growth media

The complete growth medium for culturing of LNCaP cells consisted of RPMI 1640 media supplemented with 10% FBS. The complete growth medium for PC3 and HeLa cultures was a 1:1 mixture of Dulbecco's modified Eagle's medium and F12 medium (DMEM-F12) supplemented with 10% FBS.

2.6 Cell culture seeding

Cryotubes of frozen cells for each cell line that had been stored in liquid nitrogen were quickly thawed in a 37ºC water-bath. Thawed cells were then re-suspended in 4mL of warm growth medium, centrifuged for five minutes at 500 rpm, and the supernatant was discarded to remove the remaining DMSO. Cells were re-suspended in 3mL growth medium and added to a 25 mL cell culture flasks.

All cells were subsequently incubated at 37ºC and 5% CO_2 in a CO_2 incubator. The media was changed the following day.

2.7 Cell culture maintenance

All cell lines were grown at 37ºC, with 5% CO_2 in respective media. Cells were inspected using inverted microscope for morphology and bacterial infection at least once every 2 days. Media was changed every three to four days. When cultures reached 90% confluency they were passaged into a new 25mL cell culture flask and a cryotube for cryopreservation or harvested for analysis (Figure 2.3). To prevent genetic drift, cells beyond passage 20 were not used for any downstream analysis (Kwist *et al.,* 2019).

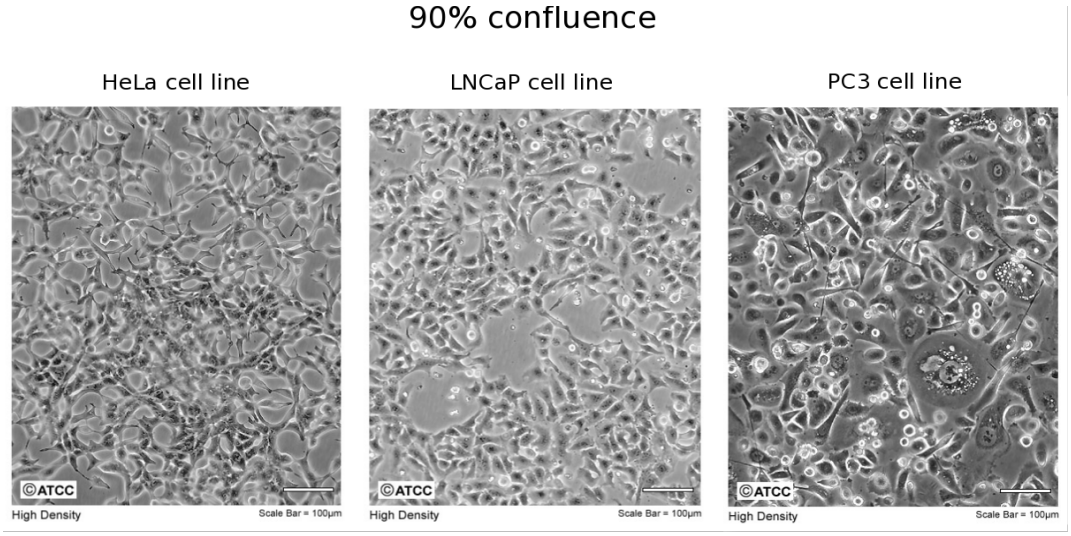

Figure 2.3 – Prostatic cell cultures

The three cell cultures under study are depicted in the high-density state (90% confluent).

Passaging was performed by first removing the media by an aspirator and cells were then washed with 1.0 mL warm phosphate-buffered saline (PBS). The PBS was then aspirated. For LNCaP, 1.0 mL of versene (0.58g EDTA:1L DPBS) was added, immediately followed by 1.0 mL EDTA Trypsin (0.05% Trypsin : 0.53mM EDTA). For other cell lines, 1.0 mL of Trypsin only (0.05%) was added. The cells were returned to the incubator to allow the Trypsin to disassociate the cells in suspension. Versene is used for LNCaP cells to assist with disassociation because this cell line is particularly agglomerated.

Following Trypsinisation, 1 mL of medium was added to the trypsin containing cells. The cell re-suspension was then transferred to a 15 mL centrifuge tube. The cells were then pelleted at 500 rpm for 5 minutes. The medium was aspirated immediately after centrifugation and the cell pellet was re-suspended in 5 mL warm complete medium. 3 mL of this cell containing medium was then pipetted into a new 2 mL cell culture flask. The remaining 2 mL of medium was either discarded or used for cell harvesting for analysis.

Cells were plated at approximately the densities detailed in Table 2.1 and checked daily for confluence and bacterial contamination. Cells tend to become unhealthy when contaminated with bacteria as this causes their mRNA profiles to deviate from that of a healthy cell, rendering the sample unsuitable for analysis.

Cell densities

Cell line	No. of cells pipetted in flask
HeLa	1.0×10^5
PC3	1.0×10^5
LNCaP	2.0×10^5

Table 2.1

Cell densities for plating cells in a cell culture flask.

2.8 Cryopreservation

Following trypsinisation and centrifugation, pelleted cells were re-suspended in room temperature (25ºC) cell culture freezing media (Appendix 5.8). The cells in freezing media were aliquoted into cryotubes (1.0 mL each) at a density of 1.0 *x* 10^6 cells per aliquot and transferred to a Nalgene Cryotube 1ºC freezing container (Nalgene Labware™) containing 100% isopropyl alcohol. This container was then placed in a -80ºC freezer. This enabled the cells to be gradually cooled to -80ºC at a rate of -1ºC $minute^{-1}$. Once cooled to -80ºC (within 2 days after being placed in the -80ºC freezer), the cryotubes were stored in liquid nitrogen.

2.9 Determination of cell concentration and viability

Cell concentration was determined using an automated cell counter (TC20™ Automated Cell Counter, Bio-rad USA) and viability was determined with trypan blue. 20µL of re-suspended cells was added to 20µL of trypan blue and incubated for 2 minutes. 10µL of this cell-trypan blue mixture was pipetted onto a slide for automated counting.

2.10 Cell harvesting and isolation

In the final step of passage, the remaining 2 mL of the 5 mL medium containing cells was pipetted into a new 15mL centrifuge tube. This cell suspension was centrifuged at 500 rpm for 5 minutes. The medium was then aspirated and the cell pellet was re-suspended in 1mL of PBS. The cells were then counted and diluted with PBS to 100 cells μL^{-1}. The cells are easier to manipulate for harvesting with a light microscope-guided mouth suction micro-pippette at this dilution.

1 mL of PBS containing 100 cells μL^{-1} was aliquoted to a 1.5 mL Eppendorf tube. 5 µL of PBS containing 100 cells μL^{-1} was transferred to a microscope slide. The light microscope magnification was set to the 40x objective. A micromanipulator was used to advance the micropipette toward the solution on the slide into the light path. When the pipette tip was touching the cell soma to be harvested, gentle suction by mouth was used until the cell entered the pipette tip. The harvested cell

or cells were then deposited into strips of 0.2 ml thin-walled PCR tubes in the lowest possible volume of H_2O (preferably ≤1 µl). The harvested cells were then stored at -80ºC.

Cells were harvested as quickly as possible while outside of the incubator. This is because the cells tend to become unhealthy when at temperatures lower than 37ºC and this causes their mRNA profiles to deviate from that of a healthy cell. All single cells and bulk cells of the same type were derived from the same passage and cell culture.

2.11 Preparation for reverse transcription

Reaction mix 1 (Table 2.2) was prepared in a 1.5 mL low bind Eppendorf tube. The Triton detergent assisted in maximizing cell lysis and RNase inhibitor assisted in protecting mRNA from RNase activity. The External RNA Controls Consortium (ERCC) control mix allows the assessment of the fidelity of the amplification method. This control mix consists of pre-formulated blends of 92 transcripts, derived and traceable from NIST-certified DNA plasmids. The transcripts are designed to be 250 to 2,000 nt in length, which mimic natural eukaryotic mRNAs. DNase I was not added due to its partial RNase activity (Woo *et al.,* 2015).

Reaction mix 1

Component	Total volume (µL)
UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen™)	87
RNase inhibitor (Enzymatics™)	5
Triton X-100 detergent (10%, Sigma-Aldrich®)	4
dNTP Mix (10mM, Invitrogen™)	100
ERCC Spike-in control mix, 1/1000 (Invitrogen™)	6

Table 2.2

Reagents added to mRNA samples to prepare for reverse transcription step.

2.5 µL of reaction mix 1 was added to each of 40 uniquely bar-coded 2.5 µL aliquots of anchored oligonucleotide dT30VN (RCD oligo-dT, 12ng/µL) in a 96-well plate. The RCD oligo-dT contained a unique 8 bp barcode, reverse transcription primer, a forward PCR primer, a T7 IVT promoter for downstream amplification and a 3' forward sequence linker (Figure 2.4). We used 40 unique barcodes that enabled us to process up to 40 samples at a time and combine them in downstream reactions and on a single sequencing lane. This feature enabled us to carry out three sequencing experiments with forty samples each.

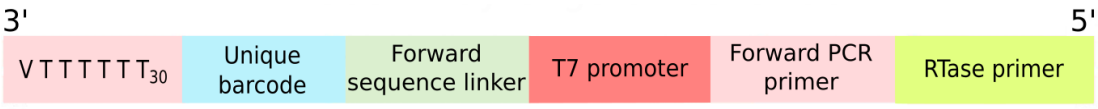

Figure 2.4 – RCD oligo-dT construct

The 30 thymidine residues at the 5' end of the construct binds to the 3' poly-adenosine tail of mature mRNA. The random residue (V) enables the construct to bind to the region where the adenosine tail is joined to the coding region.

The harvested cell aliquots were thawed after having been in -80°C conditions for at least 3 hours. The freeze/thaw resulted in partial lysis of the cell membrane and release of the mRNA for reverse transcription. RNA samples were always placed in ice and were processed without delay to avoid substantial degradation of the single cell pico-gram amounts of RNA. The 5 µL reaction mix 1 with oligo-dT was added to PCR tubes containing RNA. A multi-channel manual Biopette pipette was used to combine and mix the solution to reduce labour (Figure 2.5).

This mixture was then incubated at 72°C for 3 min and immediately put back on ice. This enabled the anchored oligo-dT primer to hybridize to the beginning of the poly-A tail of all the mRNA molecules.

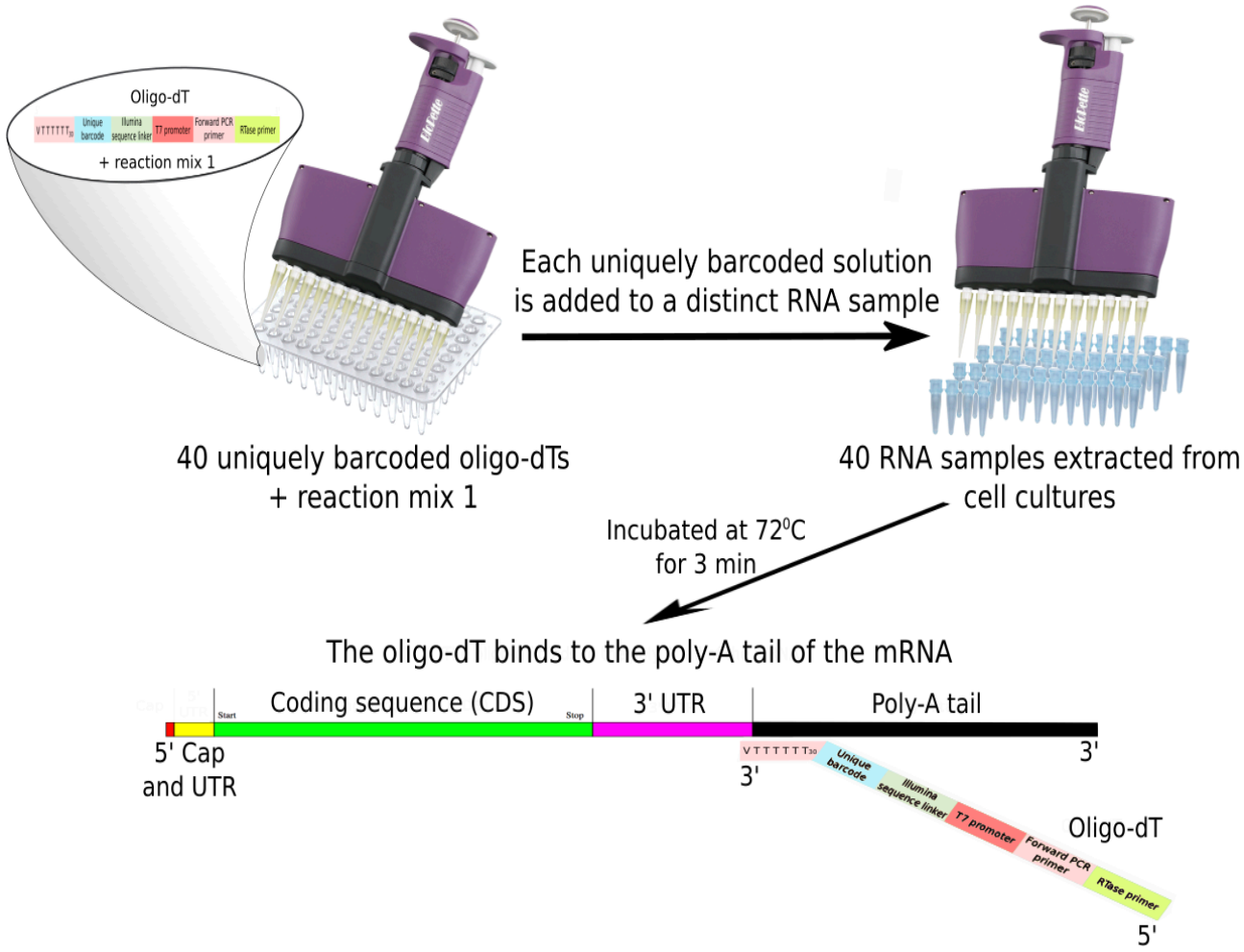

Figure 2.5 – Incorporation of the oligo-dT construct

40 uniquely barcoded RCD oligo-dTs were each mixed with reaction mix 1 and then added to 40 RNA samples. This mixture was then incubated, allowing the oligo-dT to anneal to the beginning of the poly-A tail.

2.12 Reverse transcription of mRNA to cDNA

5.7µL of reverse transcription mix 1 (RT mix 1) was added to each sample. RT mix 1 was pre-prepared for all reactions, plus one ghost sample to account for pipetting inaccuracies, by combining and mixing the reagents listed in Table 2.3.

Reverse transcription mix 1

Component	Volume per reaction (µL)
PrimeScript™ Reverse Transcriptase (200 U/µL, Takara Bio Inc.)	0.5
PrimeScript™ Buffer (5X, Takara Bio Inc.)	2.14
DTT (100mM, Sigma-Aldrich®)	0.25
Betaine (5 M, Sigma-Aldrich®)	2
Template switch oligonucleotide (TSO) containing Biotin (10µM)	0.6
$MgCl_2$ (250 M)	0.24

Table 2.3

Components of the first reverse transcription reaction.

After RT mix 1 was added to each sample, the reactions were incubated in a Applied Biosystems GeneAmp™ PCR system 2700 cycler according to the protocol in Table 2.4. Reactions were stored at 4ºC. The mRNA has now been converted to cDNA with a forward PCR primer on the 5' terminal and a reverse primer contained in the TSO incorporated on the 3' terminal. The cDNA can therefore by amplified in a PCR 'pre-amplification' step (Figure 1.6).

Reverse transcription protocol

Cycle Step	TEMP	Time	Cycles
Reverse transcription and template switching	42^0 C	90 min	1
RTase inactivation	75^0 C	15 min	1
Hold	4^0 C	-	1

Table 2.4

Protocol for the first reverse transcription reaction.

2.13 PCR pre-amplification

19µL of PCR mix 1 was added to each sample. PCR mix 1 was pre-prepared for all reactions plus one by combining and mixing the reagents listed in Table 2.5.

Polymerase chain reaction mix 1

Component	Volume per reaction (µL)
KAPA HiFi™ HotStart ReadyMix (2×)	15
Forward primer (0.5 µM)	0.5
Reverse primer (0.5µM)	0.5
UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen™)	3

Table 2.5

Components for the first polymerase chain reaction.

Amplification was then performed in an Applied Biosystems GeneAmp™ PCR system 2700 cycler according to the incubation protocol in Table 2.6. All samples were amplified with 18 cycles. The Agilent 2100 Bioanalyser High Sensitivity DNA assay

determined the yield and quality of the pre-amplified double stranded cDNA. The reactions were stored at 4ºC.

Polymerase chain reaction protocol 1

Cycle Step	TEMP	Time	Cycles
Initial denaturation	98^0 C	3 min	1
Denaturation	98^0 C	20 sec	18
Annealing	70^0 C	15 sec	18
Extension	72^0 C	6 min	18
Final extension	72^0 C	5 min	1
Hold	4^0 C	-	1

Table 2.6

Protocol for the first polymerase chain reaction.

2.14 PCR purification

The 40 PCR pre-amplified cDNA samples were combined into 5 aliquots in 1.5mL Eppendorf tubes and the volume of each aliquot was reduced to 100 µL by using a vacuum centrifuge (Express SC250EXP SpeedVac™ Concentrator System, Thermo Fisher Scientific, Inc.). The vacuum centrifuge was pre-heated to 65ºC and the samples were centrifuged for 30 minutes or until the sample volume reached 100 µL. 60µL of Ampure® XP magnetic beads (Beckman Coulter) were added to the vacuum centrifuged samples (0.6:1 ratio).

The 5 combined cDNA samples were purified following the instructions of Picelli *et al*., (2014), steps 17-22. The samples were eluted into 30µL of nuclease free H_2O in fresh 0.2mL thin-walled PCR tubes. The 5 samples can then be sequenced at this point according to the Nextera DNA Library Preparation Kit (Illumina, Inc.) protocol. However, we combined the 5 samples into a single volume and for further processing by an IVT step described in this protocol.

2.15 In vitro transcription (IVT)

The single combined PCR sample was further amplified by conducting a single round of IVT to convert the cDNA into aRNA. This was done using the Ambion™ MessageAmp™ II aRNA amplification kit, as per the kit manual (Part Number AM1751); we followed step IIF for an unmodified 40µL reaction. The IVT was incubated at 37ºC for 14 hours using an Applied Biosystems GeneAmp® PCR system 2700 cycler (Thermo Fisher Scientific, Inc.).

2µL of TURBO™ DNase (2 Units/µL, Life Technologies) was added immediately after the IVT incubation. The reaction was incubated with the TURBO™ DNase at 37ºC for 15 minutes using an Applied Biosystems GeneAmp™ PCR system 2700 cycler and then placed immediately on ice.

The RNA was fragmented with 9µL of NEBNext® Fragmentation Buffer (10X) and incubated at 94ºC for 90 seconds, then 2µL of NEBNext® 10X Fragmentation Stop Solution was added to cease the fragmentation reaction. The reaction was then brought up to 100µL with nuclease-free water, as per the Ambion™ MessageAmp™ II aRNA amplification kit manual, and placed on ice.

The IVT reaction was then cleaned up by RNeasy® MinElute® Cleanup spin columns (Catalogue no. 74204), as per the manufacturer's instructions. The sample was eluted into a 0.5 mL Eppendorf tube by 14µL of nuclease-free water that was pre-warmed to 55ºC. The aRNA sample was then stored at -20ºC.

2.16 Reverse transcription of aRNA to cDNA

The aRNA was reverse transcribed back to cDNA by adding 8µL of reverse transcription mix 2 (RT mix 2) to a 2µL aliquot of the aRNA sample. RT mix 2 was pre-prepared by combining and mixing the reagents listed in Table 2.7.

Reverse transcription mix 2

Component	Volume per reaction (µL)
PrimeScript™ Reverse Transcriptase (200 U/µL, Takara Bio Inc.)	1
Primescript™ Buffer (5X, Takara Bio Inc.)	2
Oligonucleotide (Oligo) 2*	1
UltraPure™ DNase/Rnase-Free Distilled Water (Invitrogen™)	4
Template switch oligonucleotide (TSO) containing Biotin (10µM)	0.6
$MgCl_2$ (250 M)	0.24

Table 2.7

Components of the second reverse transcription reaction.

The RT reaction was incubated according to the protocol in Table 2.8. The reaction was then placed on ice.

*Oligo 2 contains a random hexamer and a reverse PCR priming site for further downstream PCR required to add the 5' sequencing linker.

Reverse transcription protocol 2

Cycle Step	TEMP	Time	Cycles
Initial denaturation	25^0 C	5 min	1
Reverse transcription	37^0 C	30 min	1
RTase inactivation	85^0 C	1 min	1
Hold	4^0 C	-	1

Table 2.8

Protocol for the second reverse transcription reaction.

The RT product was purified by Ampure XP® magnetic beads (Beckman Coulter). 12.5 µL of the magnetic beads were added to the 10 µl RT product and it was purified following the instructions of Picelli *et al.* (2014), steps 17-22. The RT product bound to magnetic beads was eluted in 50 µL PCR mix 2 (see Table 2.9) for library amplification by a second PCR. The reverse PCR primer contained the reverse Illumina sequence linker.

Polymerase chain reaction mix 2

Component	Volume per reaction (µL)
KAPA HiFi™ HotStart ReadyMix (2×)	20
Forward primer (0.5 µM)	2.5
Reverse primer (0.5 µM)†	2.5
UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen™)	20

Table 2.9

Components for the second polymerase chain reaction.

In summation, the addition of the reverse PCR priming site during the second RT produced a library with the full complement of sites required to generate a finished library suitable for PCR paired-ended sequencing. Forward and reverse primers used during PCR incorporated 5' and 3' sequencing linkers into the PCR product. Amplification was then performed in an Applied Biosystems GeneAmp™ PCR system 2700 cycler according to the protocol in Table 2.10. The reactions were stored at 4ºC.

† The reverse primer contains the 5' sequence linker that enables paired end sequencing.

Polymerase chain reaction protocol 2

Cycle Step	TEMP	Time	Cycles
Initial denaturation	98^0 C	5 min	1
Denaturation	98^0 C	30 sec	6
Annealing	65^0 C	30 sec	6
Extension	72^0 C	15 sec	6
Final extension	72^0 C	5 min	1
Hold	4^0 C	-	1

Table 2.10

Protocol for the second polymerase chain reaction.

50µL of Ampure® XP magnetic beads were added to the 50µL RT product sample (1:1 ratio). The sample was purified following the instructions of Picelli *et al.* (2014), steps 17-22. The samples were eluted into 30µL of nuclease free water in a fresh 0.2mL thin-walled PCR tube.

The Agilent 2100 Bioanalyser High Sensitivity DNA assay determined the purified DNA yield, quality and size. The Qubit High Sensitivity DNA assay determined the DNA quantity. The sample was diluted to 4 nM for DNA sequencing.

2.17 RNA sequencing

We performed single-end 150 base length sequencing of the DNA library using an Illumina MiSeq instrument at the University of Otago, Department of Biochemistry, as per the manufacturer's instructions. The sequencing took approximately 21 hours.

The cDNA libraries were prepared for sequencing by Dr Robert Day through tagmentation using the Nextera XT library preparation kit, following the user manual instructions. We performed single-end 150 base length sequencing of the DNA library. This was done using an Illumina MiSeq instrument at the University of Otago, Department of Biochemistry, as per the manufacturer's instructions. The sequencing

took approximately 21 hours. 22-25 million reads are regularly obtained from a single lane for an Illumina MiSeq when performing single-end 150 base length sequencing. We multiplexed 40 libraries in a single lane, which means an average sequencing depth of between 550,000 and 625,000 reads can be obtained from each of the 40 libraries.

2.18 Data processing and statistical analysis

Raw sequences and quality scores in FastQ format were exported from Illumina BaseSpace for all clusters that passed filtering. The 40 libraries were de-multiplexed by the unique barcodes using the Sabre software package (https://github.com/najoshi/sabre). Linkers were removed and reads trimmed to 49 bp and any reads less than 20 bp were removed to increase read quality while maintaining specificity (Moreton, *et al*., 2014).

The read data was aligned to human genome build GrCh37 using the Bowtie2 and Cufflinks software packages. The aligned files were converted to BAM/SAM files by Bowtie 2 and imported into SeqMonk (http://www.bioinformatics.babraham.ac.uk/projects/seqmonk). SeqMonk was used to annotate mRNA splice variants (herein referred to as 'gene isoforms') and quantify reads mapping to gene isoforms. Gene isoforms making up the transcriptome were annotated using the gene or mRNA probe generator functions of SeqMonk. The single cell data from Ramsköld *et al.* (2012) was imported into SeqMonk in the BED file format from the Gene Expression Omnibus (accession number GSE38495).

SeqMonk's RNA quantitation pipeline was used for single cell sample analysis only (see section 3.2 of Chapter 3). The transcript feature option was set to 'gene'. Spliced transcripts were merged into a single representative value and corrected for DNA contamination. SeqMonk's 'mRNA feature probe generator' was used for single cell detection in a milieu (see section 3.3 of Chapter 3), without using a SeqMonk

analysis pipeline. The SeqMonk genome annotations are based on the Ensembl releases of GRCh37 in 2015.

The raw counts for samples with the same number of cells were normalized to the total read count by scaling up all the reads for each sample to the largest datastore using SeqMonk. Ramsköld *et al.* (2012) data was gene length corrected by converting expression to the RPKM unit using SeqMonk. Feature reports were generated with exactly overlapping probes for each analysis and converted to CSV files for further gene expression analysis with the R statistical programming language.

Data formatting, log2 transformations, volcano plots, MA plots, dendrogram hierarchical clustering, bootstrap calculations, heatmaps and *k*-medoid (Partitioning Around Medoids) clustering of principal components were done using R packages (R Core Team, 2017; Reynolds *et al*., 1992). Data for principal components analysis and heatmaps was further normalized by the standard deviation.

For the principal components analysis (PCA), the first two principal components that captured the most variance were clustered according to the partitioning around medoids (PAM) algorithm. Volcano plots depict the common differential expression values for each gene isoform across sample replicates. Differential gene isoform expression analysis was computed by empirical Bayes statistics using the LIMMA package in R (Smyth, 2004).

P values were calculated by unpaired, two-tailed T tests where the variances were treated as being equal and the pooled variance was used to estimate the overall variance. Two-tailed T tests were also used to calculate the 95% confidence intervals. Appendix 5.10 refers to the specific packages used and R code for each type of analysis. P values in volcano plots were converted to negative natural logarithmic (-log) values. P values were adjusted for multiple testing by controlling the false discovery rate with the Benjamini & Hochberg (1995) method.

Read coverage of gene isoforms was assessed with the 'Probe Trend Plot' function of SeqMonk. The quality of normalization was assessed by the 'Cumulative Distribution Plot' function in SeqMonk. Data for heatmaps and principle components analysis were normalized by standard deviation using the mean.

2.19 qPCR primer design

Quantitative real-time PCR (qPCR) with SYBR green dye was used to compare the level of putative gene expression of MAGED1 and PSA, relative to the Beta Actin housekeeping gene. This was conducted between the LNCaP and PC3 cell lines at a population level. This assisted in confirming the sequencing data.

The Primer3Plus software generated various combinations of primer sequences for MAGED1. The qPCR settings of Primer3Plus were used. The appropriate primer pair was selected after taking into consideration the following (Udvardi *et al.,* 2008):

1. **Primer Length**

   The optimal primer length is 18-25 bp. This length allows for specificity.

2. **Primer Melting Temperature (Tm)**

   Tm is defined as the temperature at which half of the DNA duplex will dissociate to become single stranded, and indicates duplex stability. Udvardi *et al.* (2008) suggest a Tm of 60ºC +/- 1ºC for qPCR.

3. **Primer annealing temperature**

   Normal annealing temperatures are between 50 and 65ºC. Higher annealing temperatures may produce insufficient primer template hybridization resulting in lower PCR product. Lower annealing temperatures tend to result in non-specific products. Both forward and reverse primers should have annealing temperatures that are similar.

4. **Guanine – Cytosine (G+C) content**

   The recommended G+C content of qPCR primers is between 40% and 60%.

5. **Avoid cross homology**

   A BLAST analysis of the primer sequences against the NCBI server was performed to test for specificity.

6. **Amplicon Length**

Maximum amplicon size should not exceed 400 bp (ideally 50-150 bases).

7. **Runs and Repeats**

   The probes should not have runs of identical nucleotides (especially four or more consecutive Gs), G+C content should be 30-80%, there should be more Cs than Gs, and not a G at the 5' end.

2.20 qPCR primers

Table 2.11 illustrates the qPCR primer sequences for quantifying the relative genetic expression of MAGED1 and PSA. The PSA qPCR primers are from Hessels *et al.,* 2003.

Gene name	MAGED1	PSA (KLK3)
Forward primer	5'-CCTGCCTCATCCTTTAACCA-3'	5'-AGCATTGAACCAGAGGAGTTCT-3'
Reverse primer	5'-CCAATTGTTCTTGCCATCCT-3'	5'-CCCGAGCAG GTGCTTTTG-3'
Region amplified	Nucleotides 10417-10645, inside intronic region (NCBI Reference Sequence: NG_012559.1).	Nucleotides 4024-5700, spanning exons 3, 4 and 5 (GenBank #M27274).

Table 2.11

qPCR primer sequences for analysis of MAGED1 and PSA gene expression.

2.21 RT-qPCR protocol

The harvested cell line aliquots were thawed, resulting in lysis of the cell membrane and release of the RNA for reverse transcription. RNA samples were always placed in ice. The RNA quantity was measured by the Nanodrop 1000 Spectrophotometer (Thermo Fisher Scientific Inc.). The RNA from both cell lines was diluted in the appropriate volume of nuclease-free water so that both PC3 and LNCaP RNA was present at 15 ng/µl. 15 ng of RNA from either cell line was added to RT-qPCR reaction mix 1 (see Table 2.12).

RT-qPCR reaction mix 1

Component	Total volume (µL)
UltraPure™ DNase/RNase-Free Distilled	2.5
RNase inhibitor (Enzymatics™)	1
RCD oligo-dT (2 µL)/dNTP mix (2 µL, 10mM,	1.5

Table 2.12

Reagents added to mRNA samples to prepare for RT-qPCR.

This mixture was then incubated at 72 °C for 3 min and immediately put back on ice. This enabled the anchored oligo-dT primer to hybridize to the beginning of the poly-A tail of all the mRNA molecules.

5.7µL of reverse transcription mix 3 (RT mix 3, see Table 2.13) was added to each sample. RT mix 1 was pre-prepared for all reactions plus one by combining and mixing the reagents listed in Table 2.2.

Reverse transcription mix 3

Component	Volume per reaction (µL)
PrimeScript™ Reverse Transcriptase (200	0.5
PrimeScript™ Buffer (5X, Takara Bio Inc.)	2.14
DTT (100mM, Sigma-Aldrich®)	0.25
Betaine (5 M, Sigma-Aldrich®)	2
$MgCl_2$ (250 M)	0.24

Table 2.13

Components of the reverse transcription reaction for qPCR.

After RT mix 3 was added to each sample, the reactions were incubated at 42ºC for 90 minutes and then at 70ºC for 15 minutes in a Applied Biosystems GeneAmp™ PCR system 2700 cycler. Reactions were stored at 4ºC. The mRNA has now been converted to cDNA. The 10 µL cDNA reaction is diluted in 90 µL of nuclease-free water to prevent the reverse transcription reagents interfering with the qPCR reaction.

Each sample was run in triplicate on the CFX Connect™ Real-Time PCR Detection System. Sterilized water was used as a no template control (NTC). 5.48 µL of qPCR reaction mix (see Table 2.14) was added to 4 µL of the cDNA, 4 µL of the water control and 4 µL of a reverse transcription reaction lacking a reverse transcriptase enzyme as an RT- control.

qPCR reaction mix

Component	Volume per reaction (µL)
KAPA™ SYBR® FAST qPCR Master Mix Universal (Kapa Biosystems Inc.)	5
Forward primer (20mM)	0.14
Reverse primer (20mM)	0.14
ReactionReady™ GC-Rich Reagent (SuperArray Bioscience Corp.)	0.2

Table 2.14

Components of the qPCR.

The qPCR reaction was incubated according to the protocol in Table 2.15.

qPCR protocol

Cycle Step	TEMP	Time	Cycles
Initial denaturation	95^0 C	3 min	1
Denaturation	95^0 C	15 sec	45
‡Anneal/Extension	60^0 C	1 min	45
Disassociation curve initiation	65^0 C	-	1
Annealing gradient	0.5^0 C - 95^0 C	*	-

Table 2.15

Protocol of the qPCR.

2.22 qPCR analysis

Standard curves were generated to quantify gene expression. qPCR runs were considered successful if the slope of the standard curve was between -3.10 and -3.50, the correlation coefficient was above 0.98 and triplicates were within 0.5 of a threshold cycle of each other. A disassociation curve was also run to ensure the absence of non-specific product. Bar graphs representing quantification of gene expression were generated with the ggplot2 package for R. Quantification was normalized to the Beta-actin housekeeping gene (Raff *et al.*, 1997)

2.23 Personnel contributions

Dr. Robert Day developed the method and conducted the PC3 and LNCaP in HeLa mix experiments. Tanis Godwin operated the micro-pipette to pick the cells. The author cultured all the cells and conducted the LNCaP in PC3 mix experiment while Dr. Robert Day ran this experiment on the Illumina Mi-seq instrument. The author discovered the 'hedgehog' problem on the Agilent 2100 Bioanalyser High Sensitivity

* See manufacturers instructions for the Applied Biosystems GeneAmp™ PCR system 2700 cycler.

DNA assay while Dr. Robert Day resolved this issue by attaching a biotin molecule to the 5’ end of the TSO.

Chapter 3

3. Results

The first aim of this project was to troubleshoot a method to measure global gene expression of individual cells. This chapter presents the results from troubleshooting the method and three major whole transcriptome experiments outlined in Chapter 2, section 2.2. Each of these experiments were conducted on separate Illumina MiSeq sequencing runs.

Results from troubleshooting and optimization of our methodology are presented first (section 3.1), followed by an analysis of single cell transcriptomes (section 3.2). We compared these single cell transcriptomes to single cell LNCaP and PC3 data generated by the method described by Ramsköld *et al.*, (2012).

In section 3.3 we present results on the detection of a single PC3 or LNCaP cell in an unrelated background population of cells (either 29 HeLa cells or 199 LNCaP cells). All single cell and multi-cell transcriptomes were sequenced using the method described in Chapter 2 (sections 2.2 - 2.17).

3.1 Optimal PCR pre-amplification of the cDNA library

Prior to applying the method to single cell samples, it was necessary to optimize the initial PCR pre-amplification (see methodology in section 2.12). We used the Agilent 2100 Bioanalyser High Sensitivity DNA assay to establish the optimal yield and quality of the cDNA library following the initial reverse transcription. Significant artifact noise in the cDNA library produced from the initial RT-PCR (following sections 2.9 - 2.13) had to be resolved before the method could be applied to single cells from cell culture.

3.1.1 Resolving artifacts in the cDNA library

The first major hurdle for amplifying the reverse transcribed cDNA by PCR was to remove artifacts from the PCR product. The most prominent artifact was a 'hedgehog' pattern most likely caused by concatenation of the TSO end of the cDNA (Figures 3.1A and 3.1C). Figure 3.1 demonstrates that this artifact is resolved by using a biotin molecule attached to the 5' end of the TSO (Figures 3.1B and 3.1D). This blocked artificial expansion of the 5' end of the cDNA. We compared 1ng/µl and 10pg/µl of PC3 derived RNA template. The wide library spread in trace C indicates RNA degradation (Picelli *et al*., 2013). See Figure 5.1 for the negative RNA control (Appendix 5.1).

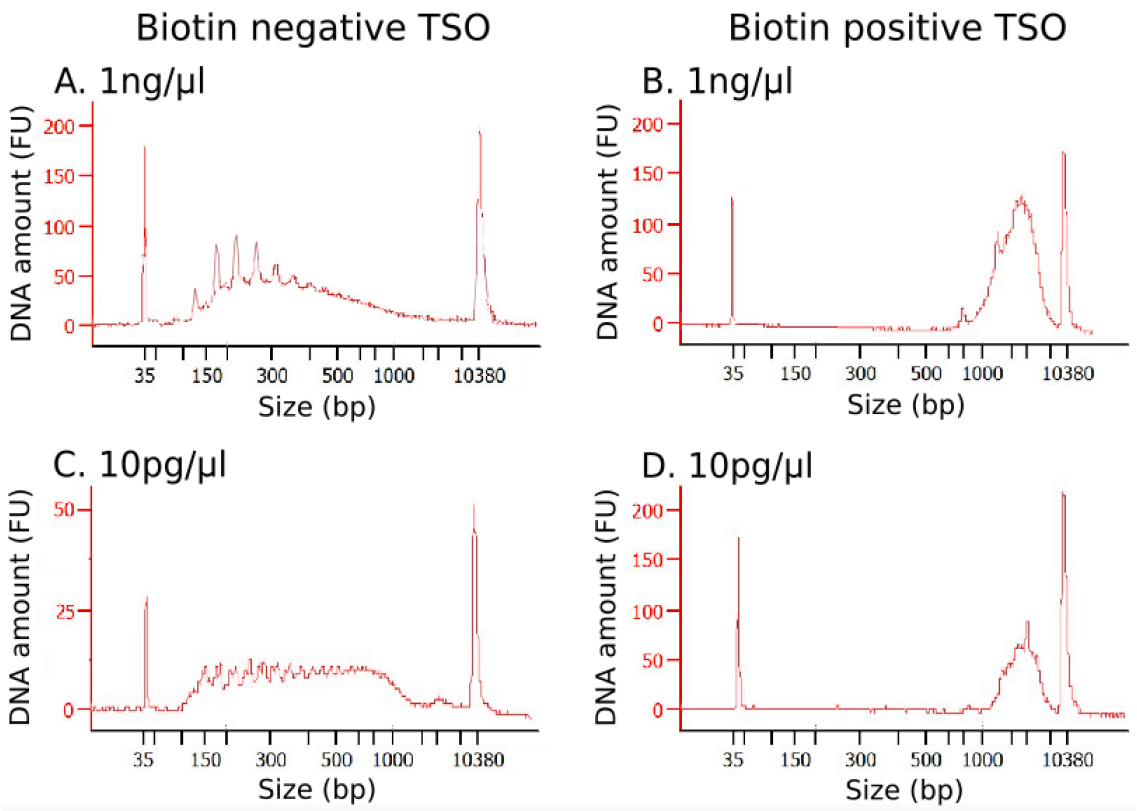

Figure 3.1 – cDNA library profile with and without the biotin TSO

Agilent 2100 Bioanalyser High Sensitivity DNA assay electropherogram traces of cDNA libraries were obtained from RT-PCR of RNA derived from lysed aggregate PC3 cells and diluted to 1ng/µl (A, B) and 10pg/µl (C, D). DNA amount is measured in fluorescence units (FU). Peaks at 35 bp and 10380 bp are gel migration markers. Traces A and C are representative examples of the characteristic 'hedgehog' pattern in samples using the biotin negative TSO. Traces B and D are representative examples of a successful library amplification using the biotin positive TSO.

3.1.2 Optimal PCR polymerase

The second consideration was ensuring the appropriate Taq polymerase was selected for optimal PCR pre-amplification. We assessed the amplification efficiency of Kapa 2G Robust polymerase and Kapa HiFi Hotstart polymerase (Kapa Biosystems). The RNA template was the HeLa RNA control from the Ambion™ MessageAmp™ II aRNA amplification kit. We assessed polymerase activity on 100ng/µl and 10ng/µl of RNA template derived from PC3 cells. Library quality and yield was assessed by examining the electropherogram produced from the Bioanalyser assay of the cDNA product (Figure 3.2).

The electropherogram profile of a successful cDNA pre-amplification is a prominent peak between ~1.5 – 2 kb (Picelli *et al.*, 2014). Kapa 2G Robust polymerase produces a

wider range of fragments, particularly under 200 bp at smaller RNA concentrations**.** This indicates undesirable primer dimers. See the Figure 5.2 for negative RNA control (Appendix 5.1).

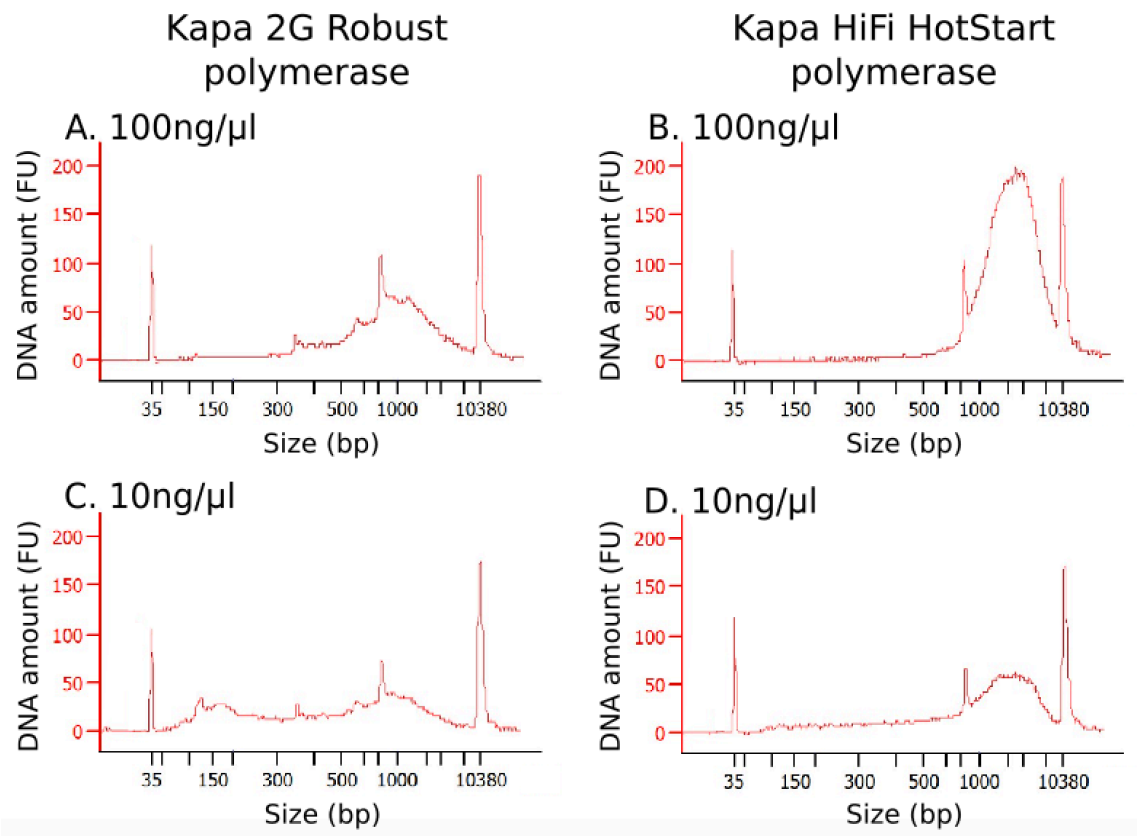

Figure 3.2 – cDNA library profile of two PCR polymerases.

Agilent 2100 Bioanalyser High Sensitivity DNA assay electropherogram traces of cDNA libraries were obtained from RT-PCR of RNA derived from HeLa RNA control and diluted to 100ng/µl (A, B) and 10ng/µl (C, D). DNA amount is measured in fluorescence units (FU). Peaks at 35 bp and 10380 bp are gel migration markers. Traces B and D are representative examples of a successful cDNA library using the Kapa HiFi HotStart polymerase. Traces A and C are representative examples of a cDNA library amplified by Kapa 2G Robust polymerase.

3.2 RNA-seq of single cells

We applied our single cell transcriptome amplification method to sixteen isolated single cells, eight from each of the PC3 and LNCaP cell lines (following sections 2.2 - 2.17). The eight single cells from each cell line were multiplexed in a single MiSeq sequencing run. Ramsköld *et al*., (2012) also quantified four PC3 and four LNCaP single cell transcriptomes with their PCR based method, dubbed 'Smart-Seq'. This PCR method developed into 'SMART-Seq2' (see section 1.4.3) by Picelli *et al.,* (2013, 2014). We compared the characteristics of our different approaches.

All libraries, including data from Ramsköld *et al.,* (2012), were quantitated with the RNA-Seq pipeline in SeqMonk using the gene probe generator. Isoforms for each gene were merged into a single measure with reads counted over exons only. This resulted in a total of 33,175 gene features that were quantified.

3.2.1 Genes detected per read

Ramsköld *et al.*'s sequencing approach used a single MiSeq run for each individual cell, so that the 20 million read capacity of the MiSeq sequencing platform was concentrated on one cell at a time. In contrast, our approach involved multiplexing 568 cells per sequencing run, eight of which (PC3 or LNCaP) were barcoded single cells in isolation. This meant the 20 million reads of the MiSeq platform was approximately 568 times more diluted in our single cell data compared to Ramsköld *et al*.'s samples. Consequently, our method detected significantly more genes per read on average ($P < 0.05$) for both cell lines than Ramsköld *et al.*'s method (Figure 3.3).

We detected 168 times more genes per read on average in single PC3 cells than Ramsköld *et al.,* and 94 times more genes per read on average in single LNCaP cells. We detected up to 33% of all genes in a single cell with only 0.005% of reads used on average by Ramsköld *et al*. (Appendix 5.5, Tables 5.1 and 5.2).

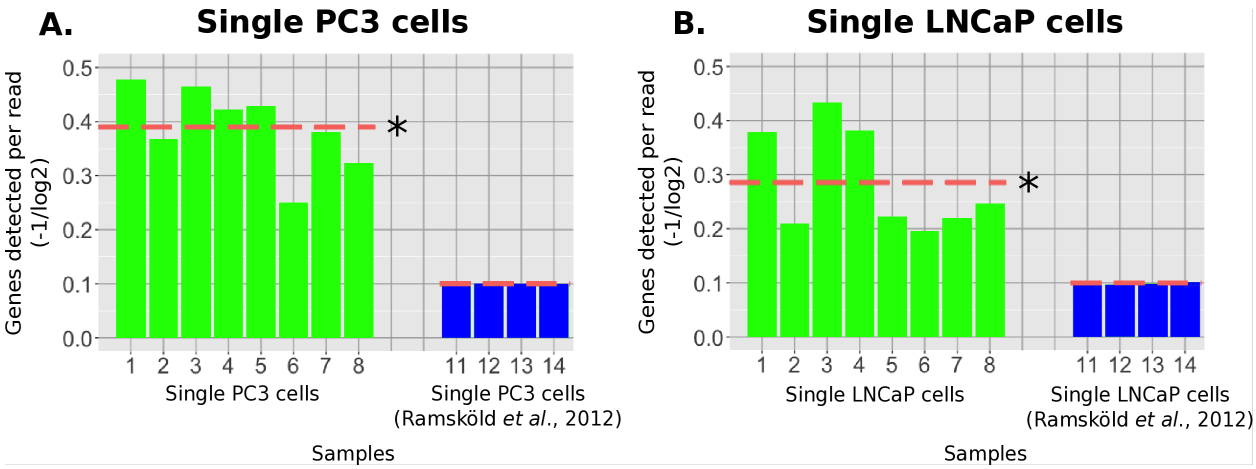

Figure 3.3 – Genes detected per read in single cell transcriptomes.

There were significantly more genes detected per read in our single PC3 (A) and LNCaP (B) cell samples (green bars) on average (red line, * = P < 0.05) compared to Ramsköld *et al.* (blue bars). The y axis was calculated by the ratio of genes detected to reads converted to the negative inverse of log2.

3.2.2 Variation across single cell transcriptomes

We quantified the degree of variability within and between single PC3 and LNCaP cell transcriptomes by calculating the non-log2 transformed Pearson correlations for Ramsköld *et al.*'s data and our data (Figure 3.4). Our single PC3 and LNCaP cells had higher median and mean Pearson correlations ($P < 0.05$) within cells of the same type, signifying more consistent transcriptomes than Ramsköld *et al*. (Figure 3.4C).

Single LNCaP cells had a relatively higher correlation than single PC3 cells within both Ramsköld *et al.*'s samples and our samples. Our single PC3 cells had an average Pearson correlation of 0.90 (CI = 0.01, n = 8) compared to the average Pearson correlation of 0.55 (CI = 0.37, n = 4) for Ramsköld *et al.*'s single PC3 cells. Our single LNCaP cells had an average Pearson correlation of 0.94 (CI = 0.01, n = 8) compared to the average Pearson correlation of 0.90 (CI = 0.03, n = 4) for Ramsköld *et al.*'s single LNCaP cells.

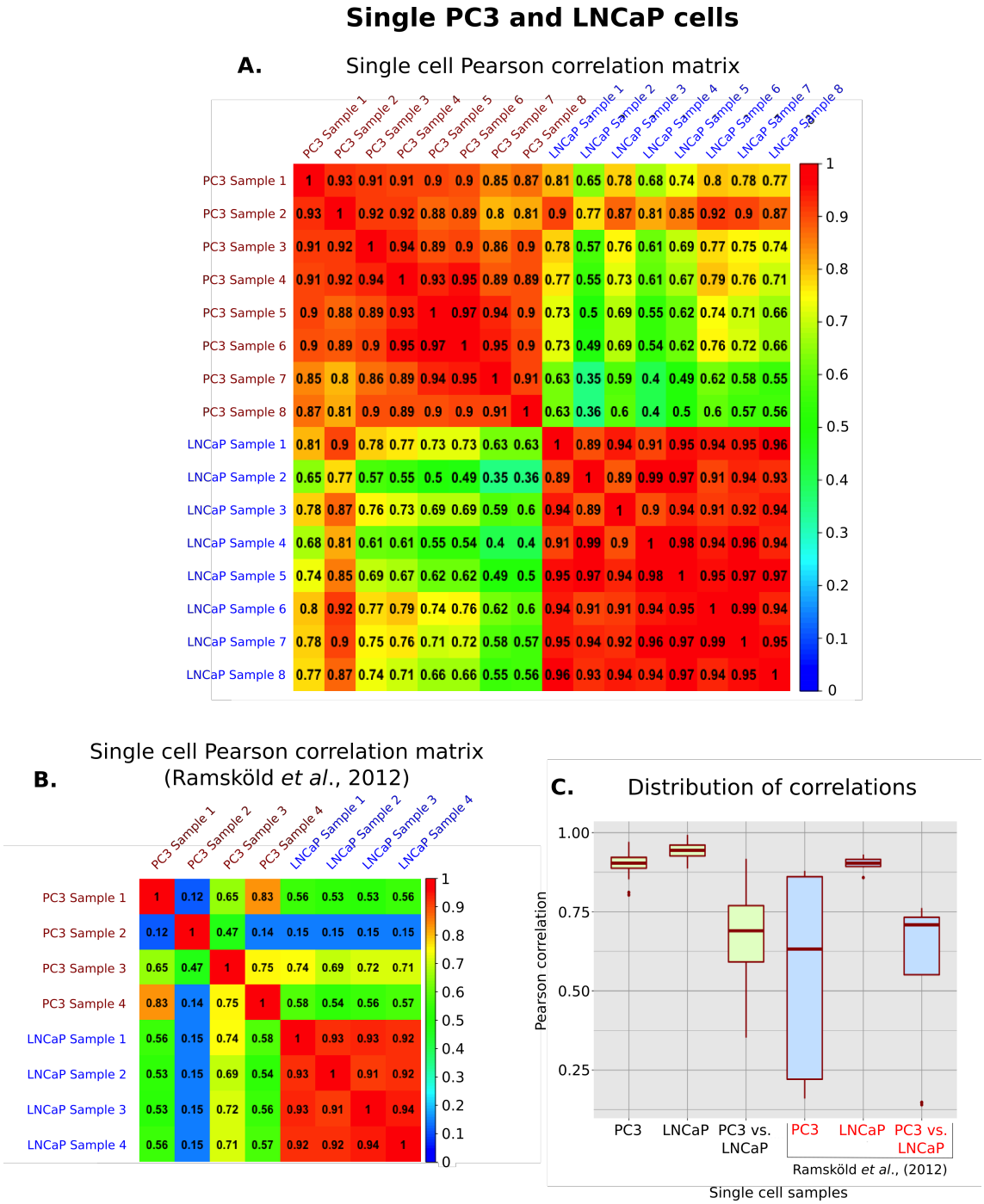

Figure 3.4 – Pearson correlations between single cells.

Pearson correlations across single cell transcriptome expression for our data (A) and Ramsköld *et al.*, (2012) (B). Figure C is a boxplot depicting the distribution of Pearson correlations for each sample set, with the median represented in the bold red. The median Pearson correlation for our single PC3 cells was 0.90 (n = 8) and for Ramsköld *et al*. 0.63 (n = 4). The median Pearson correlation for our single LNCaP cells was 0.94 (n = 8) and for Ramsköld *et al*., 0.90 (n= 4). Our median Pearson correlations were significantly higher than Ramsköld *et al*. (P < 0.05). The median correlations between PC3 cells and LNCaP cells were relatively consistent between the two methods (our data = 0.69, Ramsköld *et al.*'s = 0.71).

3.2.3 Principal components analysis of single cell transcriptomes

The transcriptome levels of single cells were analysed using principal components analysis (PCA) and the partitioning around medoids (PAM) clustering algorithm to elucidate underlying patterns. The principal component clustering demonstrates that all single cell transcriptomes clustered according to cell line of origin and the method used.

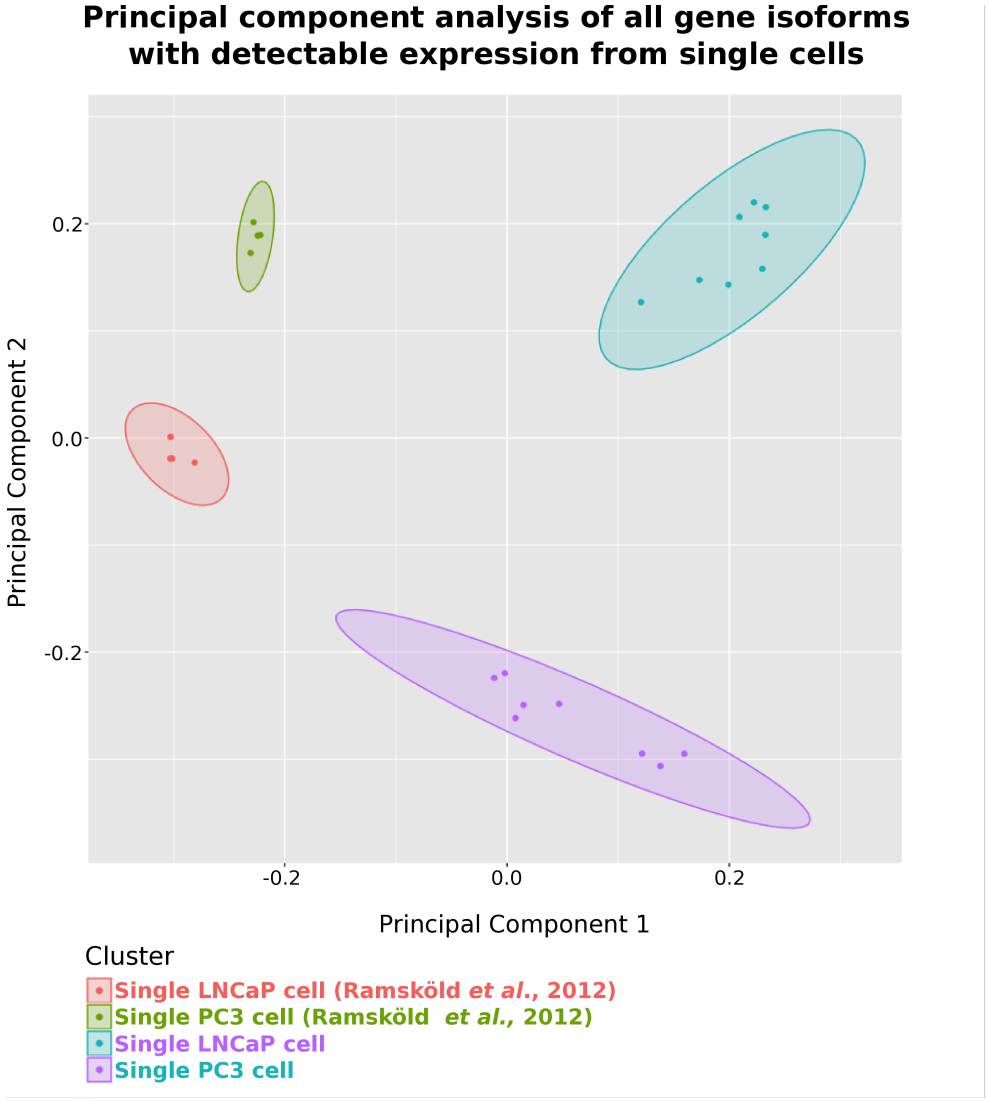

Figure 3.5 – Principal components analysis of single cells.
PCA for single LNCaP and PC3 cells from our method and Ramsköld *et al*. (2012). Data points are shown based on the first two principal components that capture the most variance and clustered according to the PAM algorithm (*k* = 4). Reads mapped to 28,854 gene features out of a total of 33,175 (86%) across all single cell samples (including Ramsköld *et al.* (2012). Ellipses assume a multivariate normal distribution with the confidence level set at 95%.

3.2.4 Read coverage across gene isoforms

As expected, our data had reads mapping almost exclusively to the 3' end of gene probe features (Figure 3.6B). In contrast, Figure 3.6A illustrates Ramskold *et al.*'s attempt at read coverage across the full length of transcripts, despite still exhibiting a 3' bias.

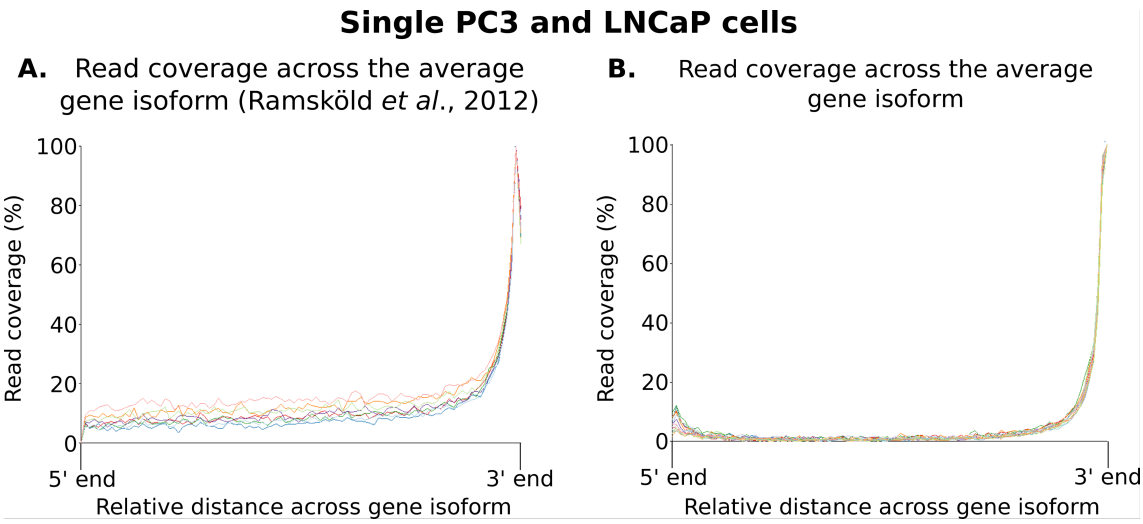

Figure 3.6 – Read coverage across gene length.

The read coverage across gene length was deliberately limited to the 3' end to simplify the quantification of gene expression for our method (B). The traces in graphs A and B represent the read coverage over the average gene isoform which is calculated by determining the number of reads overlapping with each position in each probe. These quantities were plotted out as an average over all genes by SeqMonk (see section 2.17).

3.3 Single cell detection in a background population

We applied our method to detecting single prostate cancer derived cells (from the PC3 or LNCaP cell lines) in a mileu consisting of either HeLa cells or a different type of prostate cancer cell line. We did this by first determining the mRNA splice variants (gene isoforms) over-expressed in the prostate cancer derived cell that we were trying to detect, relative to the background population sample (volcano plot A in Figure 3.7). These gene isoforms formed the basis of the signal for the single cell that we were aiming to detect.

Secondly, we identified the gene isoforms over-expressed in the sample containing a single prostate cancer derived cell in a background, relative to a pure background population sample (volcano plot B in Figure 3.7). Finally, we determined the most overly-expressed gene isoforms with a P value below 0.05 and log2 fold change >1 that were shared between volcano plots A and B. We treated the expression pattern of these shared gene isoforms as the signal for the presence of a single prostate cancer derived cell in the background population. We illustrated this signal in a heatmap (red for over-expression relative to background cells) that depicts the relative expression levels of the shared gene isoforms for individual samples (Figure 3.7).

We also included the eight isolated single cell samples from the LNCaP or PC3 cell cultures in the heatmap. If the gene isoform expression pattern of these single cells in isolation was generally consistent with the pattern of the same single cells in the background population, we could be more confident that the signal was genuine. Libraries for single cell detection were quantitated in SeqMonk using the mRNA probe generator. This allowed us to quantify mRNA splice variants of genes. This resulted in a total of 149,135 mRNA features (gene isoforms) that were quantified.

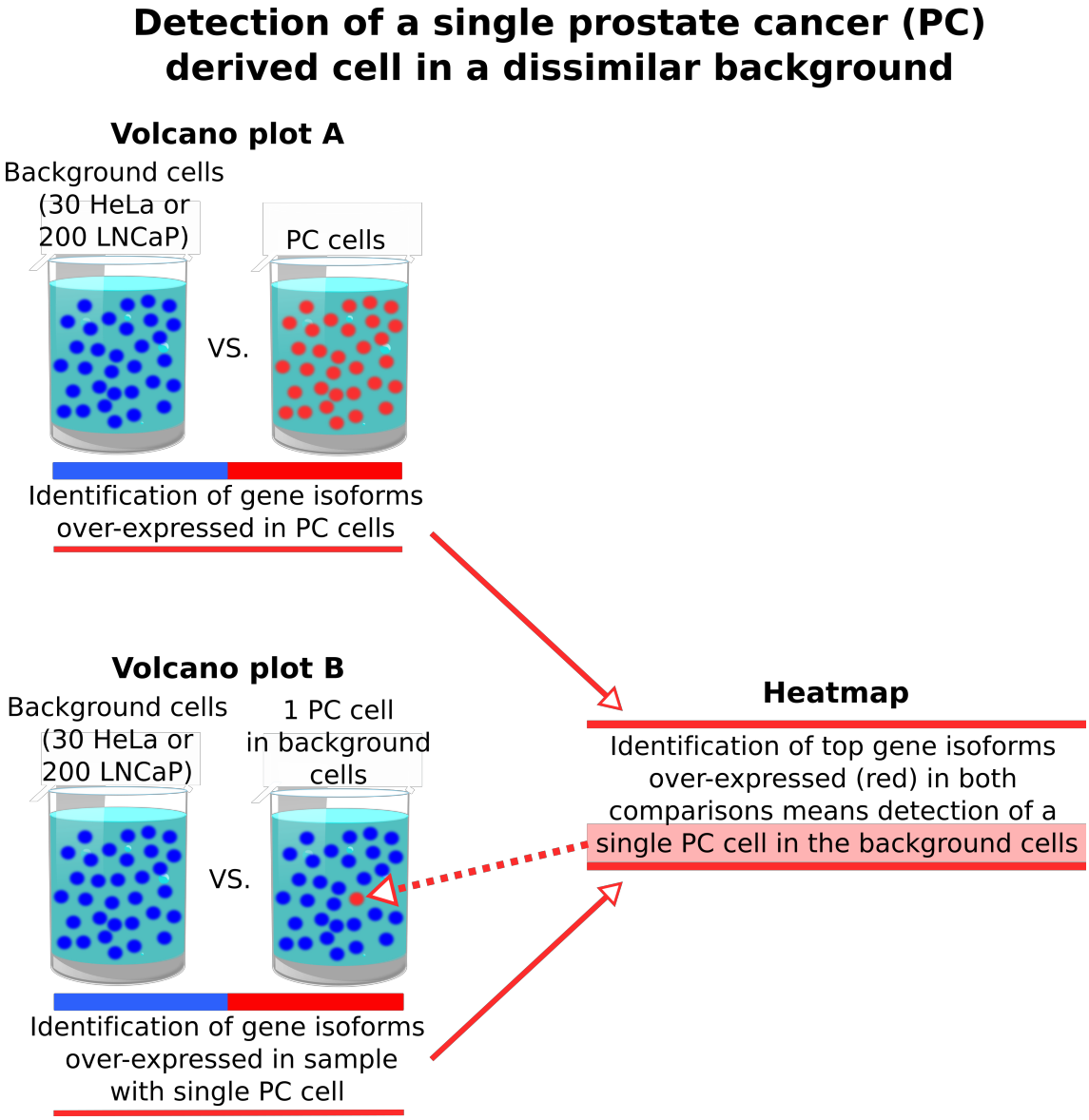

Figure 3.7 – Detection of a single prostate cancer cell in a background population.

A graphic explanation of Figures 3.8, 3.10 and 3.12. The figure illustrates the method for distinguishing a prostate cancer cell in a different cell background using differential gene isoform expression represented by two volcano plots that identify gene isoform expression specific to the prostate cancer (PC) cell of interest. The gene isoform expression pattern that is specific to the single prostate cancer cell of interest is represented by a heatmap that includes the cell in the background population, the cell of interest in isolation and the background cells alone.

3.3.1 Detection of a single PC3 cell in 29 HeLa cells

The volcano plots in Figure 3.8 illustrate the number of differentially expressed gene isoforms in PC3 cells relative to HeLa cells. We detected a single PC3 cell in 29 HeLa cells by 161 gene isoforms that were over expressed in both volcano plots A and B depicted in Figure 3.8. The most over expressed gene isoform in PC3 cells compared with HeLa cells is from MAGED1 (volcano plot A: log2 fold change = 9.3, $P < 0.01$). MAGED1 was also highly expressed in the single PC3 cell in the 29 HeLa cell background (volcano plot B: log2 fold change = 6.5, $P < 0.01$). MA plots for both volcano plots (Appendix 5.3, Figure 5.5) show slopes of the loess curves are around 1 (horizontal) at y axis = 0, indicating that there are no systemic biases requiring further normalisation (Cleveland *et al.*, 2017).

In the individual samples depicted in the heatmap in Figure 3.8, only 3 out of 8 of the single PC3 cell in the 29 HeLa background samples exhibited MAGED1 over-expression. MAGED1 was overexpressed in only 5 out of 8 single PC3 cell samples. No single gene isoform exhibited consistent over-expression in single PC3 cells that would enable detection in a HeLa background. As reported previously, PC3 samples exhibited zero PSA gene isoform expression (Tai *et al*., 2011).

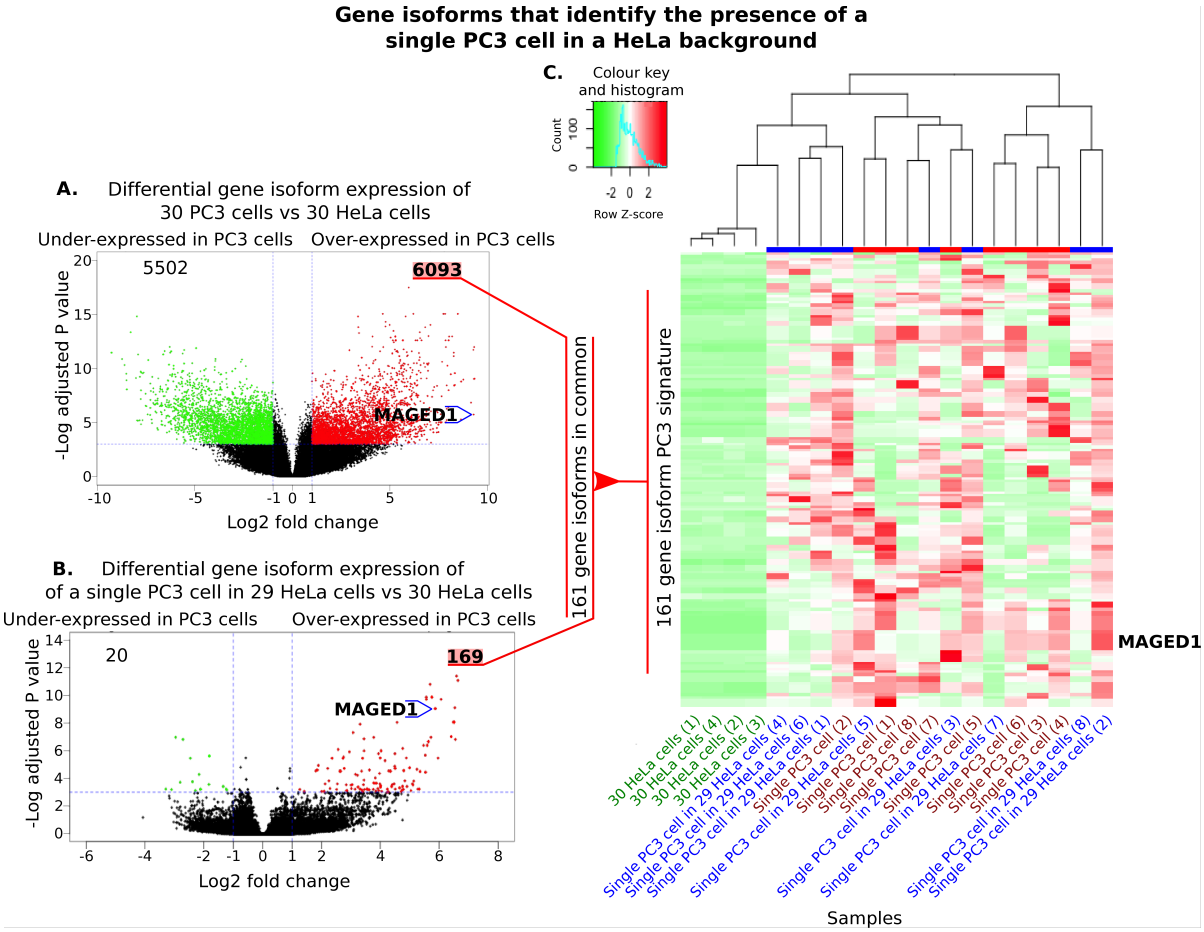

Figure 3.8 – Detection of a single PC3 cell in 29 HeLa cells.

Volcano plots of differentially expressed gene isoforms between 30 PC3 (n = 3) and HeLa cells (n = 4) (A), and a single PC3 cell in 29 HeLa cells (n = 8) and 30 PC3 cells (n = 3) (B). The highly differentially expressed (log2 fold change >1), statistically significant (P < 0.05) gene isoforms (highlighted in red) that were shared between both volcano plots were analysed in a heatmap of 30 HeLa cells, a single PC3 cell in 29 HeLa cells and single PC3 cells alone (C). 161 gene isoforms identified the presence of a single PC3 cell in 29 HeLa cells (P < 0.05). The MAGED1 gene isoform was the most over expressed gene in PC3 cells relative to HeLa cells. The 30 PC3 cells were not included in the heatmap. Sample identifier numbers are in brackets along the x axis of the heatmap.

We were also able to distinguish single PC3 cells from the 29 HeLa cell background by PCA of both the 161 gene isoform signature (Figure 3.9A) and global transcript expression levels (Figure 3.9B). PCA/PAM clustering of the 161 gene isoform signature showed that single PC3 cells in the 29 HeLa mix and single PC3 cells in isolation clustered together. PCA/PAM clustering of the global transcript levels showed that single PC3 cells in a 29 HeLa background could still be distinguished from 30 HeLa cell samples, but did not cluster with single PC3 cells in isolation.

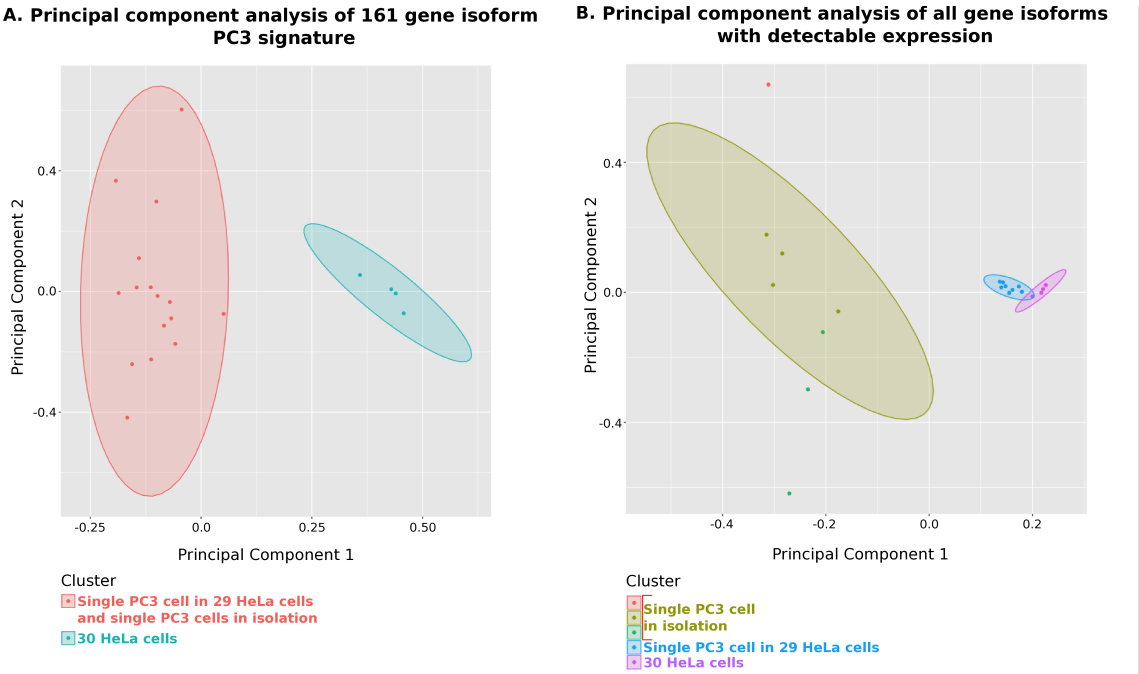

Figure 3.9 – Detection of a single PC3 cell in 29 HeLa cells (principal components analysis).

PCA/PAM clusters the single PC3 cell in 29 HeLa cells (n = 8) together with single PC3 cells in isolation (n = 8) by the 161 gene isoform signature (A, *k* = 2) but not by all detectable gene isoforms with detectable expression (B, *k* = 5). In Figure B, reads mapped to 112,845 gene isoforms out of a total of 149,135 (76%) across all samples. PAM cluster analysis distinguished PC3 cells from HeLa cells in all principal component analyses (A and B). Single PC3 cells did not cluster in a single cluster when all gene isoforms with detectable expression were analysed. Ellipses assume a multivariate normal distribution with the confidence level set at 95%.

3.3.2 Detection of a single LNCaP cell in 29 HeLa cells

The volcano plots in Figure 3.10 illustrate the number of differentially expressed gene isoforms in LNCaP cells relative to HeLa cells. We detected a single LNCaP cell in 29 HeLa cells by 64 gene isoforms that were over expressed in both volcano plots A and B depicted in Figure 3.10. The PSA gene was highlighted due to its biological and clinical relevance in prostate cancer. PSA was over-expressed in 30 LNCaP cells relative to 30 HeLa cells (volcano plot A: log2 fold change = 6.4, $P < 0.01$). The MAGED1 gene was also over expressed in 30 LNCaP cells relative to 30 HeLa cells (volcano plot A: log2 fold change = 7.1, $P < 0.05$).

PSA was over-expressed in the single LNCaP cell in 29 HeLa cell background (volcano plot B: log2 fold change = 4.0, $P < 0.05$). MAGED1 was also over expressed in the single LNCaP in 29 HeLa cell background, but not to the same degree of statistical significance (volcano plot B: log2 fold change = 4.9, $P < 0.1$). MA plots for both volcano plots (Appendix 5.3, Figure 5.6) show slopes of the loess curves are around 1 (horizontal) at y axis = 0, indicating that there are no systemic biases requiring further normalisation (Cleveland *et al.*, 2017).

In the individual samples depicted in the heatmap in Figure 3.10, only 4 out of 8 of the single LNCaP cell in the 29 HeLa background samples exhibited PSA over-expression. PSA was over-expressed in 7 out of 8 single LNCaP cell samples. Three of the single LNCaP cell in 29 HeLa cell background samples (samples 5, 6 and 7) only showed over-expression across 5 out of 64 gene isoforms in the heatmap. Sample 5 only showed weak over-expression in one gene isoform out of 64. No single gene isoform exhibited consistent over-expression in single PC3 cells that would enable detection in a HeLa background.

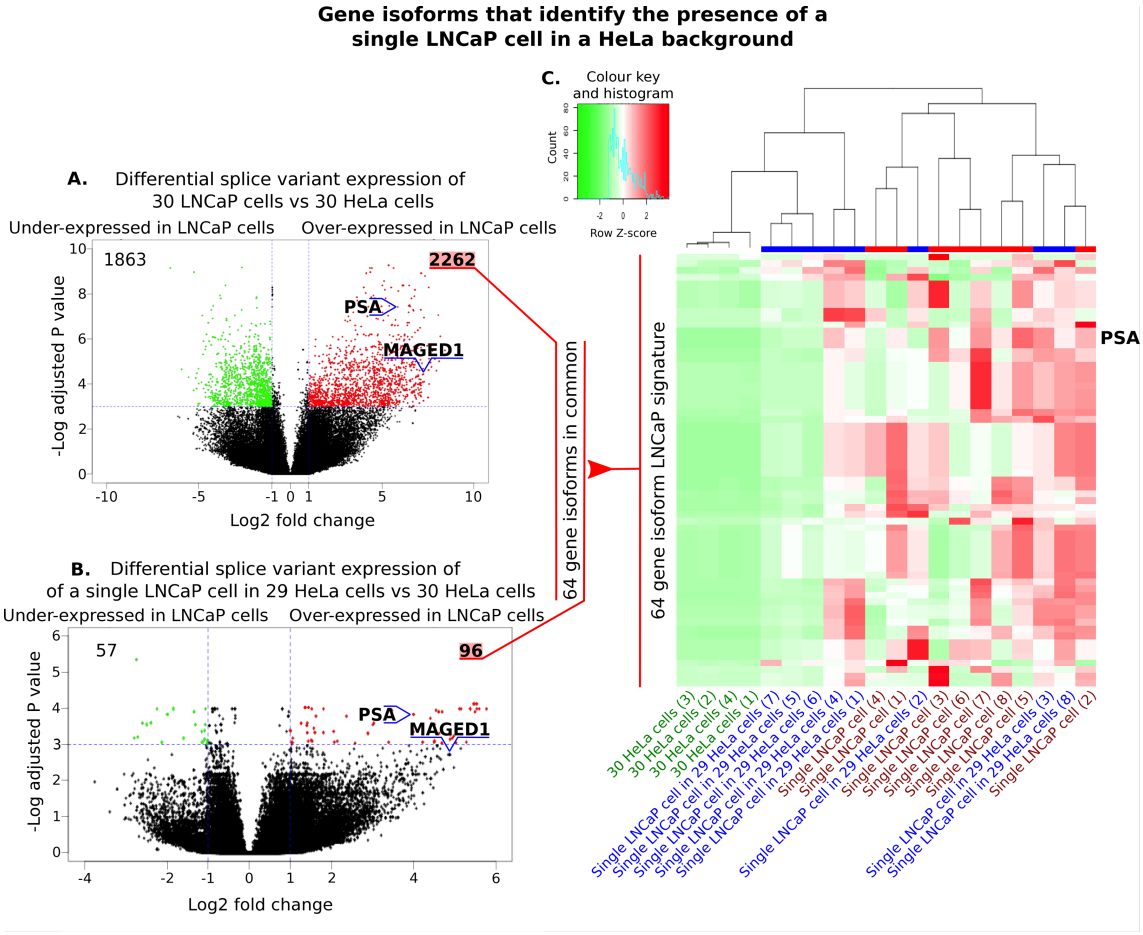

Figure 3.10 – Detection of a single LNCaP cell in 29 HeLa cells.

Volcano plots of differentially expressed gene isoforms between 30 LNCaP (n = 3) and 30 HeLa cells (n = 4) (A), and a single LNCaP cell in 29 HeLa cells (n = 8) and 30 HeLa cells (n = 4) (B). The highly differentially expressed (log2 fold change >1), statistically significant (P < 0.05) gene isoforms (highlighted in red) that were shared between both volcano plots were analysed in a heatmap of 30 HeLa cells, a single LNCaP cell in 29 HeLa cells and single PC3 cells alone (C). 64 gene isoforms identified the presence of a single LNCaP cell in 29 HeLa cells (P < 0.05). The PSA gene isoform was the most consistently over-expressed gene in LNCaP cells relative to HeLa cells. The 30 LNCaP cells were not included in the heatmap. Sample identifier numbers are in brackets along the x axis of the heatmap.

We were also able to distinguish single LNCaP cells from the 29 HeLa cell background by PCA of the 64 gene isoform signature (Figure 3.9A), but not global transcript expression levels (Figure 3.9B). PCA/PAM clustering of the 64 gene isoform signature showed that single LNCaP cells in the 29 HeLa mix and single LNCaP cells in isolation clustered together. PCA/PAM clustering of the global transcript levels showed that single PC3 cells in a 29 HeLa background could not be distinguished from 30 HeLa cell samples, regardless of how many clusters (*k*) were defined in the PAM algorithm.

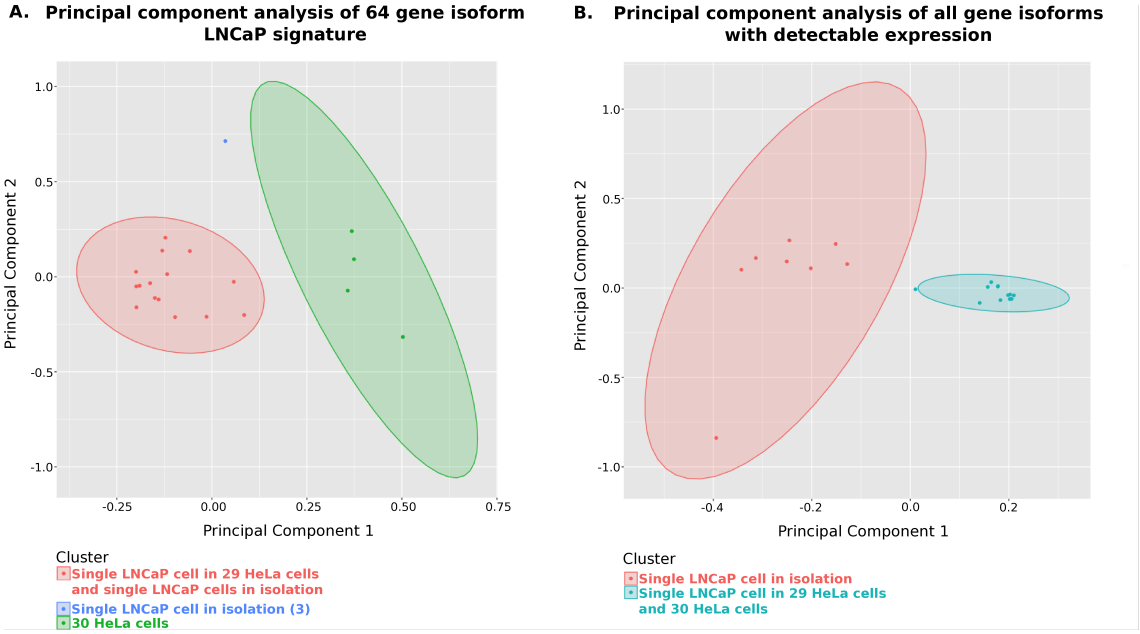

Figure 3.11 – Detection of a single LNCaP cell in 29 HeLa cells (principal components analysis).

PCA/PAM clusters the single LNCaP cell in 29 HeLa cells (n = 8) together with single LNCaP cells in isolation (n = 8) by the 64 gene isoform signature (A, *k* = 3) but not with all detectable gene isoforms (B, *k* = 2). In Figure B, reads mapped to 115,239 gene isoforms out of a total of 149,135 (77%) across all samples. PAM cluster analysis could not distinguish 30 HeLa cells (n = 4) from samples with a single LNCaP cell in 29 Hela cells when all gene isoforms with detectable expression were analysed. Ellipses assume a multivariate normal distribution with the confidence level set at 95%.

3.3.3 Detection of a single PC3 cell in 199 LNCaP cells

The volcano plots in Figure 3.12 illustrate the number of differentially expressed gene isoforms in PC3 cells relative to LNCaP cells. We detected a single PC3 cell in 199 LNCaP cells by 23 gene isoforms that were over expressed in both volcano plots A and B depicted in Figure 3.12. The gene isoform with the highest consistent over-expression in the 23 gene isoforms in the single PC3 in 199 LNCaP samples was ENTPD7 (Figure 3.12B). ENTPD7 was over-expressed in both volcano plot A (log2 fold change = 3.5, $P < 0.05$) and volcano plot B (log2 fold change = 2.7, $P < 0.05$). MA plots (Appendix 5.3, Figure 5.7) for both volcano plots show slopes of the loess curves are around 1 (horizontal) at y axis = 0, indicating that there are no systemic biases requiring further normalisation (Cleveland *et al.*, 2017).

No single gene isoform had consistent over-expression in single PC3 cells relative to LNCaP cells. In the individual samples depicted in the heatmap in Figure 3.12, only 1 out of 8 single PC3 cell samples exhibited ENTPD7 over-expression.

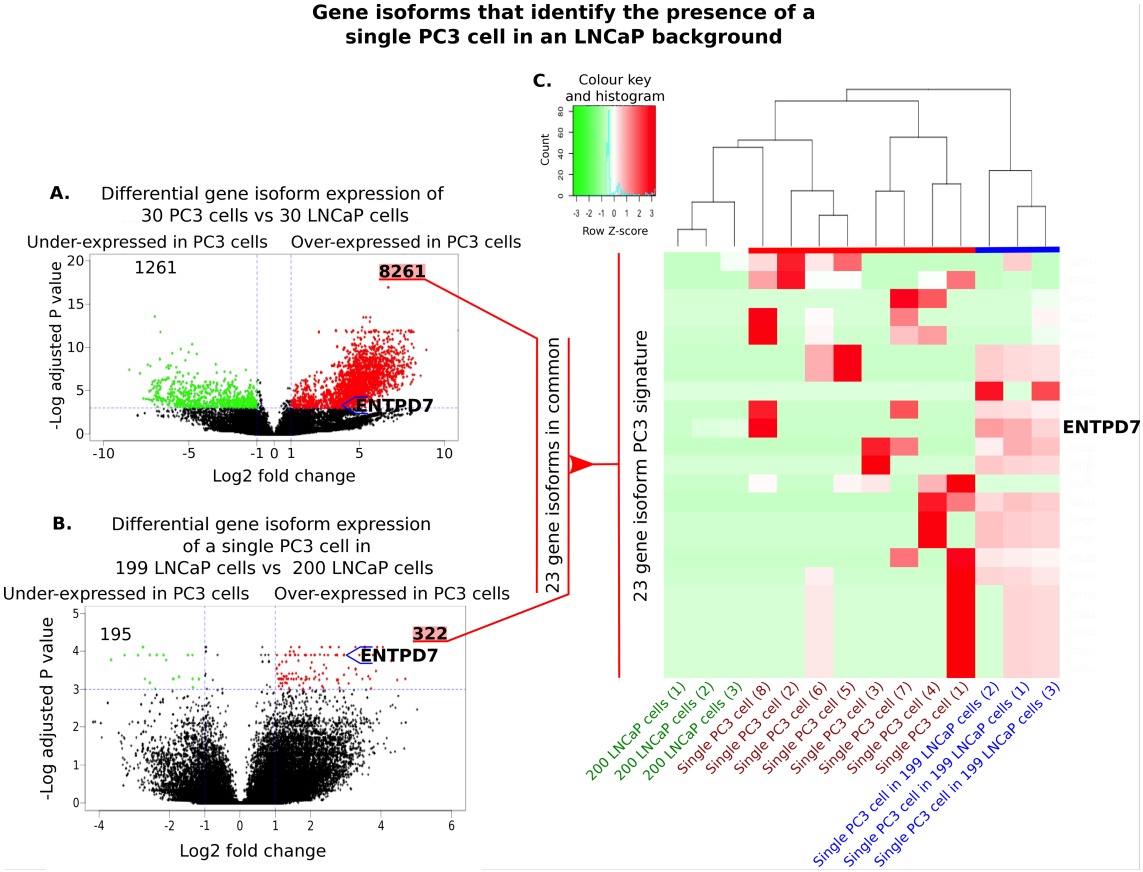

Figure 3.12 – Detection of a single PC3 cell in 199 LNCaP cells.

Volcano plots of differentially expressed gene isoforms between 30 PC3 (n = 3) and 30 LNCaP cells (n = 4) (A), and a single PC3 cell in 199 LNCaP cells (n = 3) and 200 LNCaP cells (n = 3) (B). The highly differentially expressed (log2 fold change >1), statistically significant (P < 0.05) gene isoforms (highlighted in red) that were shared between both volcano plots were analysed in a heatmap of 200 LNCaP cells, a single PC3 cell in 199 LNCaP cells and single PC3 cells alone (C). 23 gene isoforms identified the presence of a single PC3 cell in 199 LNCaP cells (P < 0.05). The ENTPD7 gene isoform was the most consistently over-expressed gene in single PC3 in 199 LNCaP cell samples relative to LNCaP cells alone. The 30 PC3 cells and 30 LNCaP cells were not included in the heatmap. Sample identifier numbers are in brackets along the x axis of the heatmap.

We could also distinguish single PC3 cells from the 199 LNCaP cell background by PCA of both the global transcript levels (Figure 3.13B) and the 23 gene isoforms identified in Figure 3.12 (Figure 3.13A). The single PC3 cells in 199 LNCaP cells grouped in a single cluster according to the expression of all detectable gene isoforms (59% of gene isoforms had detectable expression).

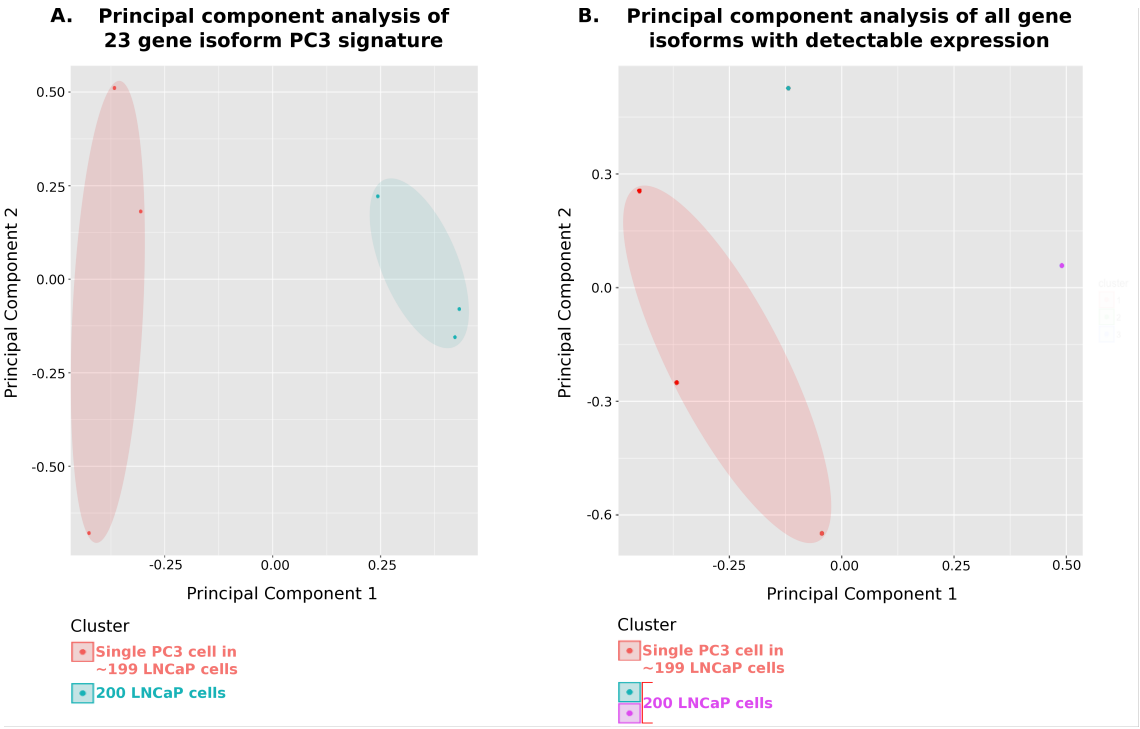

Figure 3.13 – Detection of a single PC3 cell in 199 LNCaP cells (Principal Components Analysis).

PCA/PAM of samples distinguishes the single PC3 cell in 199 LNCaP cells (n = 3) from 200 LNCaP cells (n = 3) with both the PC3 23 gene isoform signature (A, *k* = 2) all gene isoforms (B, *k* = 3). In Figure B, reads mapped to 88,083 gene isoforms out of a total of 149,135 (59%) across all samples. Ellipses only serve to highlight clusters and do not represent a multivariate normal distribution.

Single PC3 cells in isolation clustered separately from single PC3 cells in 199 LNCaP cells in both the 23 gene isoforms and global transcript levels in the PCA/PAM analysis (Appendix 5.2, Figure 5.3). There were too few points to calculate an uncertainty ellipse, so the ellipses shown in Figure 3.13 simply highlight clusters.

3.3.2 qPCR validation of RNA-seq for single cell detection

Our RNA-seq results from detecting single cells in a milieu identified high expression of the MAGED1 gene in 30 PC3 cells and single PC3 cells in 29 HeLa cells (Figure 3.8) compared to LNCaP cells (Figure 3.10). These RNA-seq results also identified PSA gene expression in LNCaP cells (Figure 3.10) but no detectable expression in PC3 cells. We validated the RNA-seq results for the expression of these specific genes with quantitative PCR (Figure 3.14).

Figures 3.14A and B illustrate PSA and MAGED1 gene expression patterns in bulk tissue samples correspond with the pattern of gene expression from RNA-seq in bulk tissue samples (Figures 3.14C and D) and single cell samples (Figures 3.14E and F).

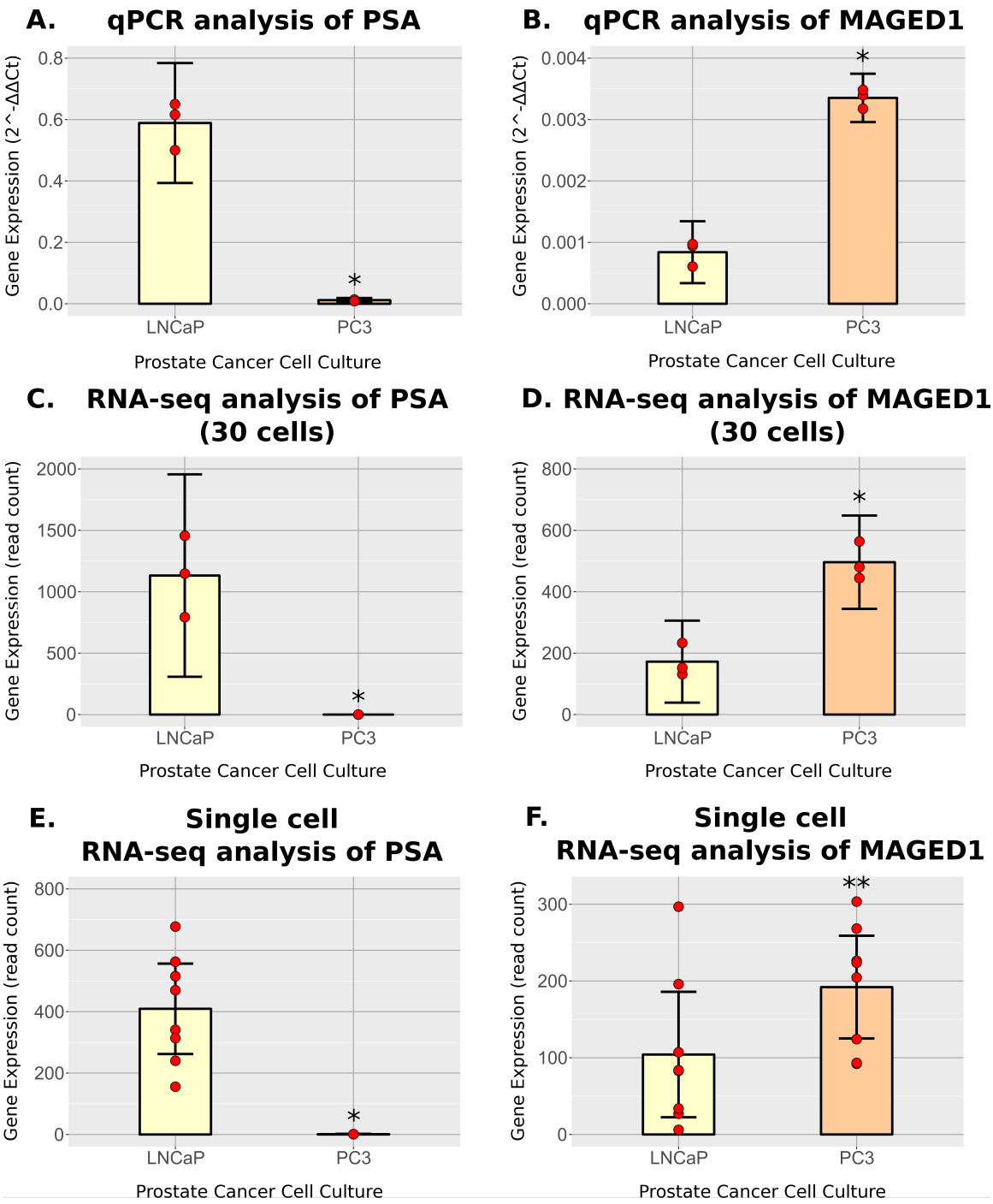

Figure 3.14 – qPCR of MAGED1 and PSA.

Gene expression analysis of the MAGED1 and PSA genes in LNCaP and PC3 cell tissues from quantitative PCR analysis of bulk cells from 100ng of RNA (A and B, triplicate samples for both cell lines) compared with RNA-seq gene expression data from 30 cells (C and D, triplicate samples for both cell lines) and single cells (E and F, eight samples for each cell line). Red data points are individual samples. Differences in expression between tissues for both genes were statistically significant (* = P>0.05, ** = P>0.1).

Chapter 4

4. Discussion

Single cell analysis technologies have undergone rapid progress in recent years in terms of numbers of processed cells. Despite this, lower throughput well based single cell technologies continue to contribute to the single cell analysis literature (Choi and Kim, 2019). These plate based methods tend to produce higher quality libraries for a lower overall cost, one of the main reasons for their continued use. It is also generally more efficient to capture rare cell types with known cell-surface markers using flow-sorting and preparing plates of single cell libraries rather than to capture more cells using a high throughput droplet based method (Griffiths *et al*., 2018). In experiments that prioritise the quality of single cell libraries over the quantity of cells, low throughput plate based methods are still preferred.

This thesis considered troubleshooting strategies for a method capable of analyzing ultra-low input RNA, including a relevant portion of a single cell transcriptome (Day *et al.,* 2018). The method is based on using PCR to pre-amplify the mRNA population, and subsequently using IVT linear amplification to produce a library amenable to analysis using the Illumina MiSeq instrument. We applied this method to analyse the single cell transcriptomes from the PC3 and LNCaP cell cultures.

The PC3 and LNCaP prostate cancer lines also have single cell transcriptome data publicly available from the PCR based study of Ramsköld *et al.*, (2012), one of the most highly cited papers in the single cell field. We did a comparative analysis of our results with Ramsköld *et al.*, (2012). The comparison buttressed our results for distinguishing single cells from the two prostate cancer cell lines. It also sheds some light on how well single cell transcriptomes replicate across studies.

We also applied this method to detect mRNA transcripts from a PC3 or LNCaP single cell that was mixed in background cells of a different type. This approach to isolating a single cell's

breadth of gene expression in a deferential background population is novel in the literature. This application has possible implications for a diagnostic test or identifying a rare sub-population with prognostic relevance in a tumour.

4.1 cDNA amplification methodology design

We utilized an oligo-dT that latches on to the poly-A tail of mRNA (Figure 1.6, Figure 2.1). This oligo-dT incorporates PCR and IVT primers, in addition to a unique barcode that allows pooling of samples for more efficient amplification (Figure 2.4). This allowed us to use a PCR amplification step (exponential amplification; mRNA to cDNA), followed by a T7 RNA polymerase driven IVT step of pooled single cell samples (linear amplification; cDNA to aRNA).

The combination of PCR and IVT remains rarely reported in single cell studies. Kurimoto *et al.* (2006) and Suslov *et al.* (2015) have pursued this approach. They both carried out a PCR 'pre-amplification', followed by a T7 RNA polymerase driven IVT. Both studies have serious shortcomings regarding transcriptome wide analysis. Suslov *et al.*'s analysis of single cells was limited to 44 genes from a microarray. The Kurimoto study claims to have produced a microarray approach, despite not presenting any data in this regard.

There are several advantages to utilising PCR followed by IVT in the context of single cell transcriptome analysis. The preliminary PCR step creates double stranded cDNA, which is a prerequisite for T7 RNA polymerase activity. In our method, the primary intention of IVT is to simplify gene expression data analysis by converting the full length reads to 3' end only. T7 polymerase amplifies from the 3' end and is limited to amplifying between 200 and up to 4000bp (Bolon and Graham 2011; Skinner *et al*., 2004). A 3' enriched library makes it unnecessary to normalise gene expression to gene length.

We found that reads unexpectedly aligned to regions far beyond the poly-A tail 3' end of mRNA, including intronic regions potentially in precursor mRNA (Appendix 5.4, Figure 5.8). We hypothesise that this could be due to internal poly Adenine sequences. Nam *et al.* (2002) show that even anchored Oligo-dT potentially had a 51% rate of internal poly-A priming

(Figure 4.1). We show that the clear majority of our transcripts were limited to the 3' end (Figure 3.6).

Oligo-dT priming remains a conventional method for isolating mRNA in single cell RNA-seq studies. This is primarily because depletion of rRNA and tRNA is not possible at the single-cell level (Suslov *et al.*, 2015).

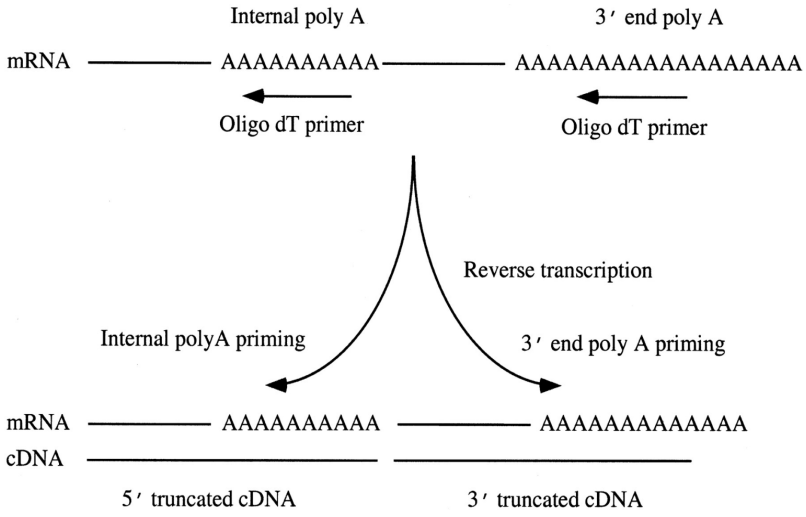

Figure 4.1 – Internal poly-A priming by the oligo-T primer
The oligo-dT can prime poly-A sequences present internally within mature mRNA or inside intronic regions in precursor mRNA in addition to the poly-A tail. The initiation of cDNA synthesis from both locations can result in two truncated cDNAs from a single mRNA template (Nam *et al.*, 2002).

Another advantage of the PCR/IVT approach is that single cell samples can be pooled after PCR amplification because of a unique barcode incorporated into the oligo-dT (Islam *et al.*, 2011). We used this approach to pool barcoded double stranded cDNA samples for a much larger starting amount that can be further amplified by IVT linear amplification by the T7 RNA polymerase. This results in the production of several milligrams of aRNA, utilizing less T7 polymerase than would be the case without a PCR pre-amplification.

4.2 Optimisation of the cDNA library amplification methodology

The first obstacle to producing a quality cDNA library was to resolve artifacts that may vitiate the quality of the single cell RNA-seq data. The most prominent artifact was a 'hedgehog' like pattern (Figure 3.1A, C). This pattern was also observed by Picelli *et al.* (2014) (Figure 4.3). Picelli *et al*. suggested that this pattern is due to 'concatamers'. Concatamers occur

RTase reaction fails to terminate; the MMLV RTase continues to reverse transcribe the TSO as if it were a continuation of the mRNA (Figure 4.2). This process of concatenation produces fragments of various lengths that partly consist of TSO copies.

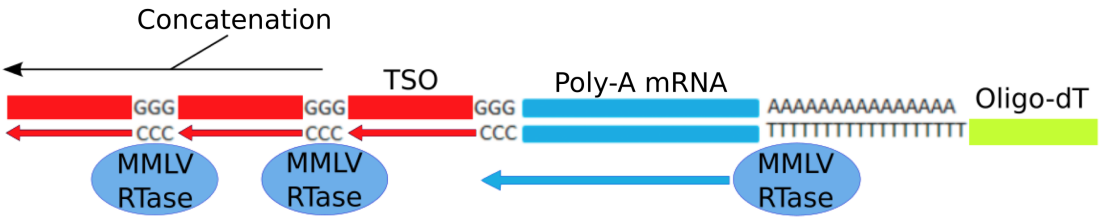

Figure 4.2 – cDNA concatenation with TSO
Model of cDNA synthesis that occurs with a TSO that lacks a biotin molecule attached to the 5' end of the TSO.

We resolve this hedgehog pattern by attaching a biotin molecule to the 5' end of the TSO which is able to terminate the MMLV RTase reaction (Figure 1B, D). Biotin modified oligonucleotides are commonly used to purify target nucleotides with streptavidin beads that bind biotin. The use of a biotin attached to the 5' end of the TSO shows that the cause of the hedgehog pattern must be concatenation.

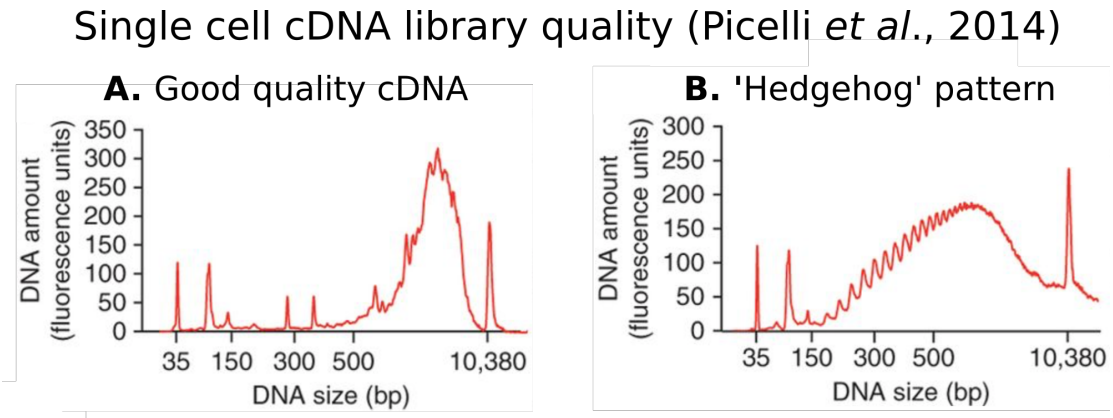

Figure 4.3 – cDNA library optimisation electropherograms
Representative examples from Picelli *et al.*, (2014) of the cDNA distributions from a successful cDNA library pre-amplification of a single mouse embryonic fibroblast cell (A) and from a 'hedgehog' pattern cDNA library pre-amplification of a single human T cell (B).

Another issue was determining the optimal RTase enzyme for single cell cDNA library production. We found that the Kapa HiFi HotStart DNA polymerase produced the cDNA library electropherogralm pattern that most resembled a good quality library compared to Kapa 2G Robust DNA polymerase (Figure 3.2). Kapa HiFi HotStart has been found to introduce the least bias in sequencing amplified bacterial genomes when compared to

unamplified samples (Quail *et al*., 2012). The electropherograms shown in Figure 3.2 are representative examples of the cDNA libraries produced by both polymerases.

4.3 Single cell statistical analysis

There are a variety of algorithms that can be used to statistically analyse a data set, each producing wildly varying results and conclusions. Therefore, the use of an algorithm requires robust justification in the context of the technical and biological parameters that produced the data. We used principal components analysis (PCA) for the study of single cell transcriptomes, a common statistical technique in the single cell literature (Lall *et al*., 2018).

PCA illustrated the similarities and differences between samples in a simple visualization (Figures 3.5, 3.9, 3.11 and 3.13). Most variation in samples is generally captured in the first two principal components (McVean, 2009). This technique is particularly powerful for simplifying the variation across datasets with thousands of genetic variables and identifying structure caused by diverse processes. We used partitioning around the mediods (PAM) to cluster samples with substantially similar gene expression profiles according to the first two principal components.

PAM is the most common technique in *k*-medoids clustering (Teschendorff and Enver, 2017). A medoid is a centrally located point in a cluster whose average dissimilarity to all the objects in a cluster is minimal. PAM finds clusters by using a pre-defined number of clusters (*k*) as an *a priori* input. Our PCA figures show clustering results after running PAM with different choices of *k,* the number of clusters. The input number of clusters was chosen based on the result that most closely resembled the number of expected clusters in the data.

PAM and the *k*-medoids technique was chosen because it is more robust than the alternative *k*-means technique (Kaufman and Rousseeuw, 1987). This is because it is less susceptible to noise and outliers, a significant factor in single cell studies.

4.4 Normalisation of single cell data

There are several crucial design factors to consider in a single cell RNA-seq analysis experiment. These factors allow us to achieve the primary objective of our single cell analysis: to distinguish individual cells based on their distinctive biology, rather than by technical noise inherent in a methodology. The major factors can be summarized as follows (Conesa *et al*., 2016):

1) Technical biases;
2) Biological biases; and
3) Read depth.

Addressing these factors allows us to potentially identify novel sub-populations. Dissecting biological and technical bias has been the foremost issue in single cell analyses. The priority is in developing a method capable distinguishing biologically or clinically relevant information in populations of single cell. This means developing a laboratory method with low technical variation. It also means employing sound computational methods to normalize single cell data sets for technical and biological confounding factors.

The type of normalization depends on the objective of the experiments. Our objective was to distinguish single cell types, including single cells seeded in a population of a different cell type. This means we must normalize for technical factors and biological factors that do not depend on cell type. This may include cell cycle effects and library size.

4.4.1 Technical biases

Libraries from single cells from PC3 and LNCaP cell lines were generated in two separate experiments and eight single cells from each cell type were run on different Illumina MiSeq sequencing runs. Technical variation can occur within an experiment and can be accounted for with technical replicates from a single biological source, like a single tissue sample or cell line. This is obviously not possible with an idiosyncratic single cell. The closest technical replicate we can come to in single cell studies is individually processing single cells from the same population source through separate parallel experiments.

Ramsköld et al. (2012), processed four single cells from both the LNCaP and PC3 cell lines through individual experiments with the Genome Analyser IIx sequencing platform (Illumina). These pseudo-technical replicates permitted an accounting of technical variation across and within experiments. Ramsköld *et al.* assessed technical variation by comparing the Pearson correlations between single cells with diluted mRNA replicate samples equivalent to single cell amounts (10pg). They found higher intra-Pearson correlations between single cells of the same type (0.75-0.85) compared with 10 pg dilution replicates (0.65-0.75).

They also compared diluted 10 pg samples with unamplified 100 ng samples from the same source. The scatter plot Pearson correlation between the 10pg and unamplified 100ng samples was 0.58. This analysis identified two populations of low abundance transcripts with distorted expression, which according to Ramsköld *et al*. represented stochastic technical losses. Despite this loss, they found that most low expressed genes could be reliably detected. A microarray experiment on picogram amounts of mRNA also found the transcriptome skewed towards high abundance transcripts, but the overall transcriptome was preserved (Kurimoto *et al.*, 2006). Ramsköld *et al*. concluded on this basis that relative gene expression levels were not significantly disrupted by technical variation.

However, there are two major problems with Ramsköld *et al.*'s analysis regarding their Pearson correlations. Firstly, they use logarithmic values to calculate correlations. Correlations between logarithmic values will always be artificially higher than correlations between raw values. The true Pearson correlation for Ramsköld et al.'s single cells is reflected in Figures 3.4B and 3.4C. Our analysis reveals real Pearson correlations for Ramsköld *et al.*'s PC3 single cells are significantly lower than reported by Ramsköld *et al*. (2012). In contrast, the Pearson correlations between single Ramsköld LNCaP cells is relatively high (0.91-0.94).

Secondly, it is not clear if the poor Pearson correlation for Ramsköld *et al.*'s PC3 cells is due to technical or biological variation. This is an inherent problem with using Pearson correlations to make conclusions about technical variation. If Pearson correlations are high, this is a good indication that there is low technical and biological variation. This is the case

with our single cell data, where we have consistently high Pearson correlations across single cells of the same type (Figure 3.4A). However, a low Pearson correlation does not differentiate between biological and technical variation *per se*.

The Ramsköld *et al.* PC3 sample with the poorest Pearson correlations is Sample 2 (Figure 3.4A, Pearson correlation range = 0.12-0.47) has a similar gene detection profile compared to the other PC3 single cells (Appendix 5.5, Table 5.1; Sample 2 = 68.7% genes detected; PC3 mean = 70.3% genes detected, n = 4) with a similar number of reads (Appendix 5.5, Table 5.1; Sample 2 = 22.8 million reads; PC3 mean = 23.2 million reads, n = 4). Interestingly, PC3 sample 2 still clusters with the other PC3 cells based on the first two principal components, while maintaining an exclusive 95% confidence ellipse vis-à-vis the LNCaP cluster (Appendix 5.2, Figure 5.4). PC3 sample 2 is mostly distinguished from the other PC3 single cell samples by the second principal component, accounting for only 32% of the variance.

The PCA demonstrates that despite the low correlation with other PC3 cells, PC3 sample 2 still maintains a PC3 principal component signature. This is a possible indication that the variation might be predominantly attributed to biology, rather than a technical fault. Technical variation would likely lead to loss of gene expression characteristic of PC3 cells. In contrast, Biological variation unrelated to cell type would still retain some gene expression characteristic of PC3 cells. Further data analysis is required to confirm that this anomalous sample's gene expression behaviour is due biological processes such as variations in cell cycle.

Chapman *et al.* (2015) suggest that using transcriptome wide correlation is a poor metric for technical reproducibility in single cells. They studied correlation between two replicates of single cell equivalent amounts derived from a pool of 100 cells and found that highly expressed genes largely determined the correlations. The top 1% most highly expressed genes accounted for 15% of the correlation. This is a challenge for single cell methods in that the clear majority of genes are expressed at the 1 – 30 transcript range, including many essential RNAs (Zenklusen *et al*., 2008).

If the overall variability of single cell gene expression is skewed towards highly expressed genes this means that subtle, yet potentially clinically relevant information is being missed or distorted by studying variability at a transcriptome level. High Pearson correlations are a crude indication of low biological variability that is clinically meaningful. However, high correlations can still give a good indication of low technical variability. We can therefore state that there is probably a relatively low degree of technical variability from the high Pearson correlations across our single cells (Figures 3.4A and 3.4C).

4.4.2 Biological biases

Cell cycle has been confounding factor in distinguishing cell types in single cell transcriptome studies (Barron and Li, 2016). A significant proportion of the transcriptome is determined by cell cycle stage. Dominguez *et al.* (2016) identified 1182 genes with cell cycle dependent expression in HeLa cells. Gene expression between single cells can vary simply because of the stage at which they are sampled in the cell cycle, complicating attempts to understand biological variation between individual cells.

In addition to periodic gene expression, the cell cycle gene profile is different for different cell types. Breast tumours are notable for the range of 'mitotic traits' that correlate with prognosis, whereas kidney tumours have a homogenous cell cycle profile (Dominguez *et al*., 2016). Despite this heterogeneity in cell cycle, Dominguez *et al.* (2016) identified a core set of 67 cell cycle genes that have highly periodic expression in common with five diverse cell lines. Elevated expression of these genes is associated with aggressive breast, lung and ovarian cancers.

The heterogeneity in cell cycle genes across cell types is a major complicating factor in efforts to normalise single cells for cell cycle (Buettner *et al*. 2015; McDavid et al., 2016). This is especially so when the cell cycle genes for the tissue under study are unknown. This is the case for PC3 and LNCaP cells.

While cell cycle can potentially confound clustering of different single cell types, the degree to which cell cycle masks true physiological differences between single cells is a matter of

debate in the literature. Buettner *et al*.'s (2015) computational analysis was the first attempt to determine the degree to which cell cycle is responsible for single cell variability.

Buettner *et al*. estimated the proportion of variance attributable to a latent variable by inferring a cell-to-cell covariance matrix from the gene expression profile of 892 cell cycle genes. They determined that cell cycle accounted for more than 30% of the variation between 81 single T helper 2 cells.

McDavid *et al.* (2016) disputed the effect of cell cycle genes by citing their own analysis of single cells. They found that cell cycle only accounted for 17% of the variance in cell cycle genes and 5% in non-cell cycle genes. McDavid studied 119 cell cycle genes that passed their quality control from 930 cells across three cell lines.

McDavid *et al.* also re-analysed Buettner *et al*.'s data and explored other factors that could be responsible for the observed variation between single cells. They found that the first principal component correlated strongly (R^2>0.99) with library size. McDavid *et al.* argue that Buettner *et al*.'s latent variable is actually quantifying the variance attributable to library size, rather than cell cycle *per se*. Library size distinguishes cell cycle in Buettner *et al*.'s data just as well as the first principal component. This may be because cell size and RNA library size correlate with cell cycle (Padovan-Merhar *et al*., 2015).

McDavid *et al*. points out that a cell cycle effect remains even after correcting for the Buettner *et al.*'s latent variable. However, the variance in cell cycle corrected data for Buettner *et al*.'s T cells was still dominated by T cell differentiation, rather than cell cycle. McDavid *et al.* calculated the sums of squares to find the real variance attributable to cell cycle in Buettner *et al.*'s mESC data is actually less than 7%.

To further complicate matters, Buettner *et al.* show that 44% of all genes studied (n=2881) showed significant correlation with at least one cell-cycle gene ($P < 0.05$), Bonferroni adjusted). This means that simply removing the suite of cell cycle genes before further analysis is insufficient for removing the cell cycle effect. This has led to the development of several Bayesian unsupervised clustering techniques to remove cell cycle effects (Lee *et al.*

2018). Some of these techniques incorporate library size as a known factor for normalisation.

Buettner *et al.*'s and McDavid *et al.*'s debates show that the contribution of cell cycle to overall cell-to-cell variance can vary greatly between cell types as well as different methods of analysis. This means that it is problematic to use a conventional set of cell cycle genes for defining the cell cycle effect across cell types.

An alternative method of correcting for cell cycle might be to correct for library size, a substantiated covariate. Buettner *et al.*'s analysis shows that the contribution of cell cycle to overall variance is reduced from 30% to 5% or less when the data has been normalised for library size (Figure 4.4).

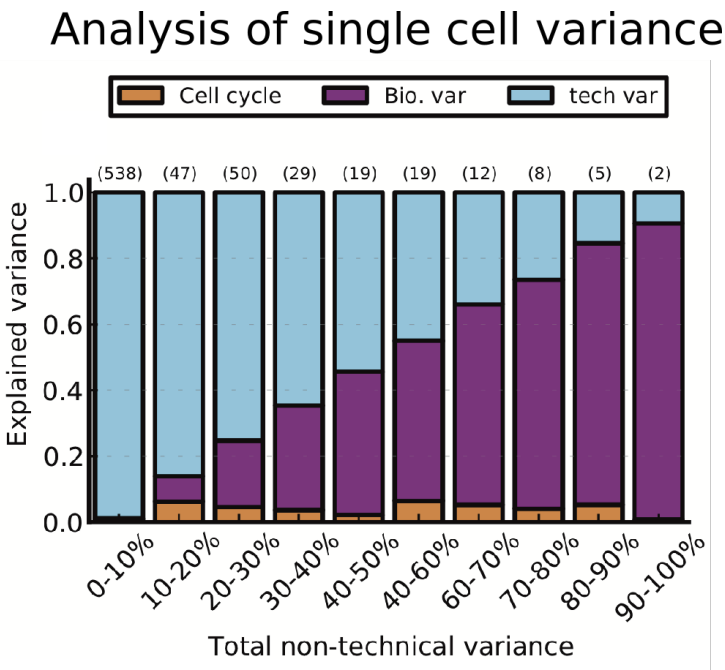

Figure 4.4 – Variance decomposition of single T helper 2 cells

Estimated contributions of cell cycle to variance between single T helper 2 cells after normalisation of library size (Buettner *et al.*, 2015).

Buettner *et al.* normalised for library size by the method described in Brennecke *et al.* (2013). Brennecke *et al.* calculated two normalisation constants, one each for technical and biological variation. Dividing the read counts by these constants brought them on to a common scale that allows comparison across samples. The technical normalisation constant accounted for differences in read depth. The biological normalisation constant was

calculated with the variation in starting RNA in each single cell. Brennecke *et al.* extrapolated the starting RNA amount in single cell samples from the read counts that mapped to the reference genome and assumed a linear relationship between starting RNA amount and the proportion of mapped reads.

We normalised our data for library size by scaling up sample total read counts to the single cell sample with the highest total read count. Obviously, this adjustment alone does not correct for cell cycle variability. In addition to this size correction, we used an empirical Bayesian method to conduct differential gene expression analysis to detect single cells in a background population. This was done with the Limma package in the R statistical platform (Phipson *et al*., 2016). Our empirical Bayesian approach leverages information from the entire dataset when making inference about each individual gene. This mitigates variances within samples of the same condition.

Phipson *et al.* (2016) use an empirical Bayesian approach to shrink the estimated sample variances for each gene towards a pooled estimate. This is accomplished by fitting a separate t-distribution for each gene across samples in the same condition. Extra degrees of freedom are added to reflect the information borrowed from global gene expression behaviour for inference about each individual gene.

In a classical Bayesian approach, prior external information is used to inform the predictive model. For example, this prior might be knowledge of the influence of cell cycle on variance. However, in this empirical Bayesian method, the prior information is informed by the marginal distribution of the data itself.

For example, single cells with less starting RNA content consistent with the early S phase in the cell cycle will tend to have more variable read count data in the final library (Ramsköld *et al*., 2012). This may be due to technical limits in sensitivity, or elevated biological stochasticity in smaller cells. This systemic variation will be reflected in that sample's marginal distribution. The Limma R package incorporates this information into the linear model fitted to each gene across all samples within each condition.

We suggest that this empirical Bayes approach to differential expression analysis mitigates the effect of cell cycle on variance across samples. This is supported by the absence of a bias trend in the loess fit in MA plots related to average gene expression plotted against differential expression to identify single cells in a background population (Appendix 5.3, Figures 5.5, 5.6, and 5.7).

In the gene signatures that differentiate a single cell from a background population, none of the single cell genes are part of the core 67 cell cycle genes (Appendix 5.5, Table A3). However, there remains significant variability in our ability to distinguish single cells in a background. For example, in the 64 gene isoform signature for detecting a single LNCaP cell in 29 HeLa cells, sample 5 only had one gene isoform showing elevated expression indicative of a single LNCaP cell (Figure 3.10). This may be due stochastic biological and technical factors, rather than systemic biases such as cell cycle.

As noted, cell cycle effects are not isolated to cell cycle genes because of gene-to-gene interactions (Buettner *et al*., 2015). Also, the core 67 cell cycle genes may not be relevant to LNCaP or PC3 cells. The absence of core 67 cell cycle genes is only an indication that cell cycle variation is not a major confounding factor in the single cell gene expression signature.

The weak LNCaP signature in LNCaP sample 5 might be due to technical variation, or the influence of cell cycle variance that has not been completely mitigated by the empirical Bayesian method. Further analysis is required to comprehensively measure the latent variability due to cell cycle in LNCaP and PC3 single cells, perhaps by unsupervised gene clustering analysis.

4.4.3 Read depth

An appropriate read depth for single cell studies has been thoroughly discussed in the literature and it is generally considered that less than a million reads is acceptable (Ramsköld *et al*., 2012). It has been demonstrated that 20,000 reads per cell is sufficient to distinguish different cell types in the same splenic tissue (Jaitin *et al*., 2014). Our lowest number of reads for a single cell was 23,462 (Appendix 5.5, Table 5.2) but this was sufficient for distinguishing cell type (Figure 3.5).

Deeper sequencing, such as attempted by Ramskold *et al.* (2014) with over 20 million reads per cell, does little to resolve differences between single cells. This is because single cells contain a tiny number of individual mRNA molecules (100-300,000 transcripts) and only a small proportion are reverse-transcribed to cDNA for analysis (Jaitin *et al*., 2014; Marinov *et al*., 2014). A possible advantage of having such extreme saturation of reads is the fine discrimination of allele specific expression.

Raw read counts can be converted to 'reads per kilobase million' (RPKM), 'fragments per kilobase million' (FPKM) or 'transcripts per million' (TPM). These units remove the bias of different gene lengths within samples and variability in sequencing depth between samples. In our experiments, normalising for read depth is particularly important because our single cell samples vary considerably in the sum of reads in the final library for analysis (LNCaP single cell read sum sample range = 24,102 – 374,678; PC3 single cell read sum sample range = 23,462 – 145,815; Appendix 5.5, Table 5.2).

It has been a common approach in the literature to normalise different read counts across samples by RPKM or FPKM. Ramsköld *et al.* (2012) use RPKM in their analysis of single PC3 and LNCaP cells to account for variability in gene length. We used the same unit for their data in our analysis for the sake of comparability with their conclusions.

The major flaw of using RPKM or FPKM units in the place of raw read counts is that the presence of highly expressed genes in a sample can disproportionately alter RPKM and FPKM values. That is because RPKM and FPKM are calculated in a way that results in different sums of reads across samples. FPKM and RPKM use different proportionality constants for each sample because the proportion of reads tend to be different across samples in different conditions. This makes the comparison of gene expression between samples using RPKM or FPKM flawed.

It is more appropriate to use TPM when comparing samples across different experimental conditions (e.g. different cell types, or different RNA-seq runs) because it uses a universal

constant. The sums of raw read counts for each sample across all conditions are divided by the same number (i.e. 'per million', the universal constant), producing a scaling factor for each sample. Read counts for each gene can be divided by this scaling factor, enabling direct comparison of relative gene expression across conditions.

In our single cell data, we found scaling up read counts in each sample to the sample with the highest read depth sufficient to normalise for read depth. This is illustrated in overlapping sample profiles in the cumulative distribution plot (Appendix 5.7, Figure 5.9). This plot demonstrates that single cell samples were well normalized for comparison. Normalisation is also confirmed by loess profiles in MA plots for the bulk samples (see Appendix 5.3, Figures 5.6 - 5.7).

The MA and cumulative distribution plots also show no significant bias from gene length. This is because our method predominantly quantified reads from the 3' end of mRNA transcripts (Figure 3.6). This meant that normalisation for gene length was not required, avoiding the need for a scaling factor.

4.5 Sensitivity and fidelity

Dissecting biological and technical bias has been the foremost issue in single cell analyses. The priority is in developing a method capable distinguishing biologically or clinically relevant information in populations of single cell. This requires a method with good sensitivity and fidelity of mRNA detection.

Sensitivity refers to a method's ability to capture and convert a specific mRNA molecule to a cDNA read in the final library analysis. The method must be able to capture a biologically meaningful number of mRNA molecules that are distributed across genes in an amount reflective of the actual gene expression.

We compared our single cell transcriptomes with single cell transcriptome data produced by a PCR based method developed by Ramsköld *et al*. (2012). We calculated the number of genes detected per read in the Illumina MiSeq libraries. The number of genes detected per

read informs us how many genes we can detect with a given number of reads. This gives an indication of the sensitivity of detection for gene expression.

Ramsköld *et al.* over-sequenced samples to detect as many genes as they could by using one cell in a single Illumina MiSeq run (~25 million reads per sample), whereas we multiplexed multiple single cells in a single run. We determined that our method detected vastly more genes per transcript compared to Ramsköld *et al.* (Figure 3.3). However, this was due to Ramsköld *et al.* using the maximum number of reads for a single cell and our protocol avoiding over-sequencing of longer genes by enriching for the 3' end. This comparison still provides a reference point indicating that our method has good sensitivity. Our method detected up to a third of the genes in single cells with only 0.005% of the reads used on average by Ramsköld *et al.,* while enabling us to clearly distinguish single cell types (Figure 3.5; Appendix 5.5, Table 5.2).

This thesis did not comprehensively evaluate the sensitivity of the methods. This would require a spike in of pre-determined amounts of artificial mRNA molecules into the sample prior to amplification. For example, ERCC spike ins are 92 poly-adenylated transcripts with a wide range of lengths and GC contents that mimic natural eukaryotic RNA (Jiang *et al.*, 2011). They are aliquoted in predetermined amounts with RNA samples under study and act as an external standard by which to measure the ability of a method to capture the endogenous mRNA.

Day *et al.* (2018) used ERCC spike ins to evaluate the sensitivity of the thesis method. They demonstrate that our method can detect approximately 48% of input mRNA molecules from a single cell. Islam *et al.* (2014) also found an average capture efficiency of 48%. Islam *et al.* also used a TSO based PCR method but included his multiplex barcode in the TSO. We included our multiplex barcode in the oligo-dT rather than the TSO to reduce the TSO's length. A shorter TSO has been shown to improve mRNA capture efficiency (Zajac *et al*., 2013).

We did not evaluate the fidelity of the methods. A method with good fidelity produces a result that is consistent with the original distribution of reads across the transcriptome, so

that the detection of expression from one gene isn't disproportionately represented in the final analysis. This may be particularly relevant to PCR based exponential amplification methods, where small initial biases can quickly escalate. One way to better ensure that the final library is faithful to the original transcriptome is to use unique molecular identifiers that attach to each mRNA molecule before amplification (Islam *et al*., 2014).

Several normalization methods have also been developed to correct for the artificially unequal distribution of sequencing reads. For example, two-parameter generialized Poisson models have been developed that simultaneously consider read depth and sequencing bias as independent parameters (Srivastava and Chen, 2010). More complex Bayesian normalization methods have been developed to detect and account for hidden gene expression determinants across multiple samples (Stegle *et al*., 2012; Mostafavi *et al*., 2013). Further study could employ these techniques to evaluate amplification fidelity in our method.

4.6 Detection of a single cell in a background population

We detected a single cell gene expression signature for either single PC3 or LNCaP cells in a background population ranging from 30 cells to 200 cells. This was done across three separate sequencing runs on the Illumina MiSeq platform:

i. Detecting single PC3 cells in a background population consisting of 29 HeLa cells;
ii. Detecting single LNCaP cells them in a background population consisting of 29 HeLa cells; and
iii. Detecting single PC3 cells in a background population consisting of 199 LNCaP cells.

The detection of single cell gene expression signatures illustrates the potential of the method to detect rare cell types in a tissue. We selected a HeLa population background because they have a very different genetic origin and profile compared to prostatic cancer cell lines. This increased the chance of success of our proof of principle experiments. HeLa cells are also readily available and easy to culture.

PC3 cultures have a modal chromosome number at 62 and LNCaP cultures range from 76 to 91 chromosomes (Ohnuki *et al*., 1980; Chu *et al*., 1983). HeLa cells have 76 to 80 heavily mutated chromosomes, with catastrophic chromothripsis across several chromosomes (Landry *et al*., 2013). Chromothripsis is a phenomenon involving karyotype level random chromosome shattering and rearrangement. This occurs in 2-3% of all cancers. PC3 and LNCaP cultures in contrast have much more localised rearrangements (Wu *et al*., 2012).

4.6.1 Detection of a single PC3 cell in 29 HeLa cells

Our method could identify single cells in all three background population experiments. We show significant gene expression variability within the signature across single cell samples (Figures 3.8, 3.10 and 3.12). Due to the ability of empirical Bayes to smooth out systemic bias we suggest that this variability is due to stochasticity in technical and biological variation, rather than say cell cycle. This is supported by the lack of core 67 cell cycle genes in the distinguishing gene signature (Appendix 5.6, Table 5.3).

We discovered a 161-gene isoform signature distinguished a single PC3 cell from a 29 HeLa cell background (Figure 3.8). The single PC3 cells in a HeLa background cluster with single PC3 cells in isolation based on this 161-gene isoform signature in the PCA (Figure 3.9A). Single PC3 cell in a HeLa background samples clustered separately from 30 HeLa cell samples according to the whole transcriptome PCA (Figure 3.9B). This demonstrates that our method (Figure 3.7) can distinguish PC3 single cells in a background based on the whole transcriptome and identify the PC3 gene expression of the single cell.

4.6.2 Detection of a single LNCaP cell in 29 HeLa cells

We identified a 64-gene isoform signature distinguished a single LNCaP cell from a 29 HeLa cell background (Figure 3.10). According to PCA, the ‘single LNCaP cells in a HeLa background’ cluster with single LNCaP cells in isolation based on this 64-gene isoform signature (Figure 3.9A). Single LNCaP cell in a HeLa background samples clustered with 30 HeLa cell samples relative to single LNCaP cells in isolation according to the whole transcriptome PCA (Figure 3.9B).

This result demonstrates that our method (Figure 3.7) can distinguish LNCaP single cells in a background based on the LNCaP signature. However, the 'single LNCaP cell in HeLa background' samples were not able to be distinguished from 30 HeLa cells according to the whole transcriptome. This might be because the single LNCaP cells have lower gene detection per read than the single PC3 cells (Figure 3.3), indicating lower sensitivity in gene detection in the LNCaP cell sequencing run than the PC3 cell sequencing run.

4.6.3 Detection of a single PC3 cell in 199 LNCaP cells

In our third experiment, we went further by increasing the size of the background population to 199 cells to determine the sensitivity of single cell detection. We also used a prostate cancer cell line as a background population (LNCaP) rather than HeLa cells to try to identify a single cell from another prostate cancer culture (PC3).

We successfully isolated a 23 gene isoform signature that identified a single PC3 cell in a 199 LNCaP cell background (Figure 3.12). We could also distinguish single PC3 cells from the 199 LNCaP cell background by PCA of both the global transcript levels (Figure 3.13B) and the 23 gene isoforms identified in Figure 3.12 (Figure 3.13A). This illustrates the potential of our method to identify rare cell types in a tissue or tumour. However, the bootstrap value for the node separating the single cell in 199 LNCaP cells from 200 LNCaP cells was 30 (approximately unbiased p-value = 97, Appendix 5.8, Figure 5.10). This low bootstrap value is reflective of the low number of gene isoforms, and indicates that the number of background cells in this experiment is close to the detection limit for individual cells.

4.7 Single cell transcriptome variability

Across all single cells detected in a background population experiments there was significant variation in gene expression. This was exemplified in the MAGED1 gene in single PC3 cells and PSA gene in single LNCaP cells. For example, only 3 out of 8 of the single PC3 cell in the 29 HeLa background samples exhibited MAGED1 over-expression (Figure 3.8). MAGED1 was over-expressed in only 5 out of 8 single PC3 cell samples.

In LNCaP single cells in a 29 HeLa background only half the samples exhibited PSA over-expression. PSA was more consistently expressed in single cells in isolation, with 7 out of 8 single LNCaP cells showing PSA over-expression (Figure 3.10). MAGED1 had lower fold change expression in LNCaP single cells vis-à-vis HeLa cells than single PC3 cells. In both experiments with single PC3 and LNCaP cell detection in a HeLa background, there was no single gene that was consistently over-expressed across all samples that would indicate the presence of a single cell.

The general PSA and MAGED1 gene expression patterns results are confirmed in the qPCR of bulk PC3 and LNCaP RNA (Figure 3.14). PSA is strongly expressed in LNCaP relative to PC3 according to qPCR results, and has no expression in PC3 tissue. This is expected because PC3 cells lack androgen receptors which are required for PSA expression (Kim and Coetzee, 2004). Increased MAGED1 expression in PC3 cells compared to LNCaP cells is also reflected in the qPCR of bulk samples (Figure 3.14B).

Despite the variability of MAGED1 expression in single cells, MAGED1 could be a reliable candidate mRNA biomarker for low abundance prostate cancer cells in urine. This presents a subject for further study. Over-expression of MAGED1 mRNA in prostate cancer biopsy tissue relative to healthy prostate tissue has been characterised previously (Kumar *et al*., 2011).

In the single PC3 cell detected in 199 LNCaP cell samples we observed 11 out of 23 gene isoforms with consistent over-expression indicating the presence of a single cell. However, in this experiment there were only three samples with a single PC3 cell. There were also more gene isoforms that distinguished PC3 cells from LNCaP cells (8261 gene isoforms, Figure 3.12) than distinguished HeLa cells from either PC3 cells (6093 gene isoforms, Figure 3.8) or LNCaP cells (2262 gene isoforms, Figure 3.10). These factors may explain a more stable gene expression signature for single PC3 cells in an LNCaP background.

The variation within the gene signature identifying a single cell in a background population is indicative of either technical or biological confounding factors. The systemic confounding effects are mitigated by the shrinkage effect of empirical Bayes (Appendix 5.3, Figures 5.5-B, 5.6-B and 5.7-B). Systemic biological effects include cell cycle and cell size. Systemic technical

effects may originate during sample processing steps, such as the conditions of mRNA capture from a single cell. Variation in pipetting accuracy may also introduce systemic biases between sample conditions.

Systemic biological (cell cycle) and technical (batch effects) variation are mitigated by empirical Bayes to a degree, however stochastic biological and technical variation will not be mitigated. Marinov *et al*. (2014) identify a significant combination of biological and technical stochastic noise in single cell gene expression. Biological stochasticity can be caused by stochasticity in alternative splicing machinery (Melamud and Moult, 2009).

Transcriptional 'bursting' is another source of gene expression stochasticity (Kaufmann and van Oudenaarden, 2007). The concept of 'bursting' revolves around the observation that genes can spontaneously fluctuate between active and repressed states of expression (Bahar *et al*., 2015). Bacterial populations use this stochasticity of gene expression to maintain expression variation that improves adaptability in the face of pathogens (Moyed and Bertrand, 1983). Cancer tumour populations have also been suggested to use this 'bet-hedging' strategy to adapt to chemotherapy and other existential threats (Sharma *et al.*, 2010).

Wide variation in gene expression in single cell gene signatures is likely to be impacted by non-systematic biological and technical stochasticity because systemic effects have been mitigated. The stochastic nature of this variation makes it more difficult to statistically model and normalise.

4.8 Conclusions and future directions

We successfully removed a 'hedge-hog' concatenation artefact in our cDNA library. This was a pre-requisite for any further analysis of our single cell method. We achieved this by attaching a biotin molecule to the 5' end of the TSO, which prevented concatenation. We further optimised the PCR protocol by selecting the best quality library produced by the Kapa HiFi HotStart DNA polymerase compared to the Kapa 2G Robust DNA polymerase.

Our high Pearson correlations within single cells of the same type illustrate that our method shows low technical variability within a single MiSeq Illumina sequencing run. According to the Pearson correlations, our method produced more reproducible results across single cells of the same type compared to Ramsköld *et al*., (2012). Our method was more sensitive than Ramsköld et al.'s method in terms of genes detected per read. However, technical replicates of single cells of the same type in separate sequencing runs would further validate these conclusions.

We conducted trouble shooting on a single cell/ultra-low input RNA method capable of high throughput gene expression analysis, producing an RNA-seq library for several thousand genes from a single prostate cancer cell. We successfully applied this method to detecting a single prostate cancer cell in HeLa and LNCaP background populations. We identified MAGED1 as a transcript that is significantly upregulated in half the PC3 single cells relative to a HeLa background and is also over-expressed in single LNCaP cells but not to the same level of statistical significance. This makes MAGED1 mRNA a candidate biomarker for low abundance prostate cancer cells in urine.

We acknowledge the limitation of the experimental design for detecting rare prostate cancer cells in a micro-population with regards to an analogous clinical context. For example, HeLa cells are not a good model of the heterogenous background of non-cancerous urinary cells. However, the experimental design might be applied to case-control analysis of ultra-low input RNA from patient urinary samples. Differential high throughput gene expression of urinary cell micro-populations between low risk patients and a patient cohort with increased risk of prostate cancer may discover RNA transcripts from intact rare pre-cancerous prostate cells. This may yield improvements in diagnostic specificity and sensitivity and evaluation of future risk.

We observed significant variation between single cells detected in background populations which we attribute to non-systematic biological and technical confounding factors. The loess profile in the MA plots of differential expression analysis for single cells in background populations show that empirical Bayes and other normalisation measures removed systemic biases, such as cell cycle. This is supported by the absence of cell cycle genes in single cell

signatures. However, biological and technical stochastic effects remain, causing considerable variation between single cells. Accurately delineating biological and technical effects across single cell samples will require further analysis.

While single cell analysis technologies will continue to push the boundaries of throughput, inexpensive and sensitive lower throughput well-based technologies will continue to remain useful for various ultra-low RNA input applications. For example, Clare *et al.* (2018) revealed new molecular characteristics that differentiate two neuronal sub-types by utilising the well based method developed by Day *et al.* (2018) and used in this thesis. Such well based single cell/ultra-low RNA input methods have shown better detection of low abundance transcripts compared to high throughput methods (Wang *et al*., 2019). Unlike high throughput microfluidic methods, well based methods are not restricted by cell size and shape. Finally, we have illustrated that well based methods dynamically accommodate a variety of experimental designs that would be impossible to implement in more rigid high throughput microfluidic technologies.

5. Appendix

5.1 Electropherogram negative controls

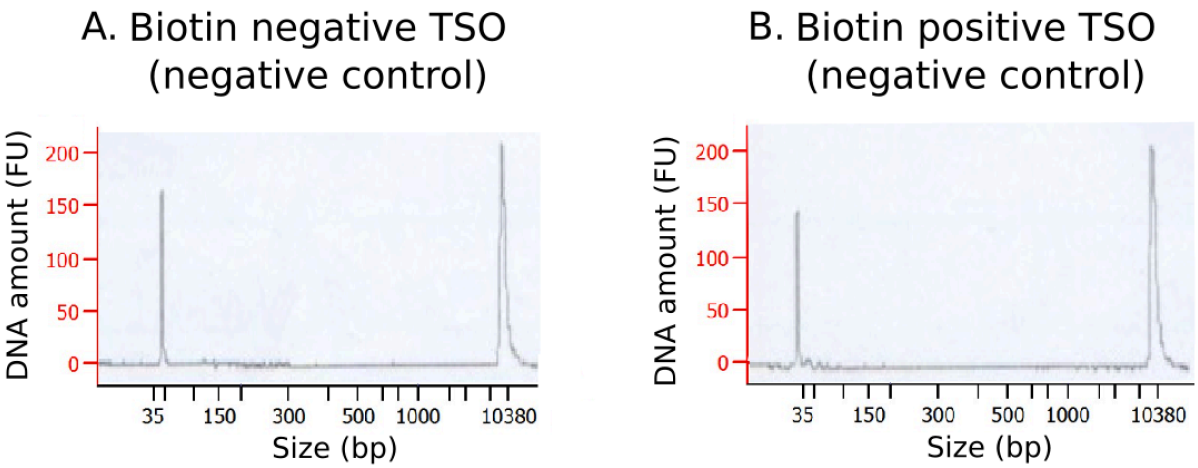

Figure 5.1 – Negative RNA controls for biotin TSO modification.

Agilent 2100 Bioanalyser High Sensitivity DNA assay electropherogram traces of cDNA libraries for negative RNA controls in biotin negative (A) and biotin positive (B) TSO. DNA amount is measured in fluorescence units (FU).

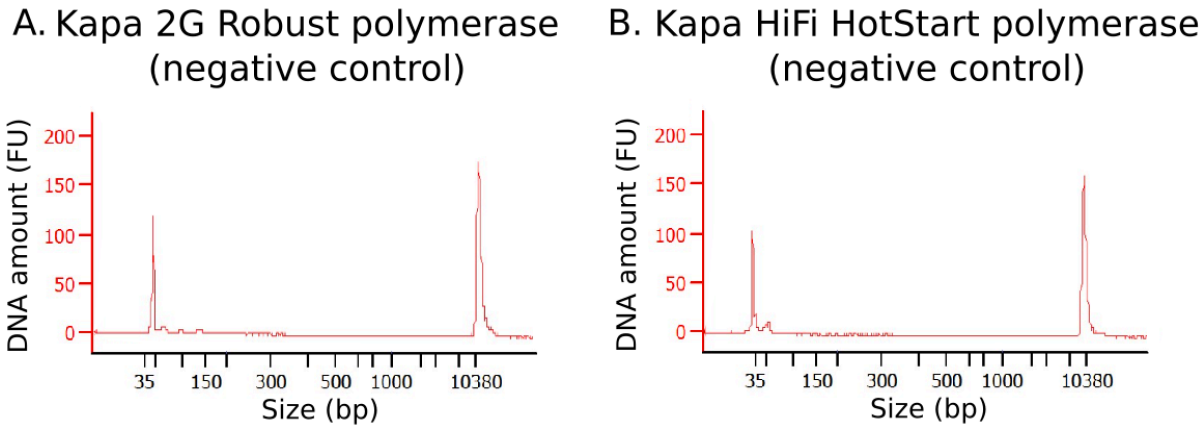

Figure 5.2 – Negative RNA controls for PCR polymerases.

Agilent 2100 Bioanalyser High Sensitivity DNA assay electropherogram traces of cDNA libraries for negative RNA controls in libraries amplified by Kapa 2G Robust polymerase (A) and Kapa HiFi HotStart polymerase (B). DNA amount is measured in fluorescence units (FU).

5.2 Principal components analysis

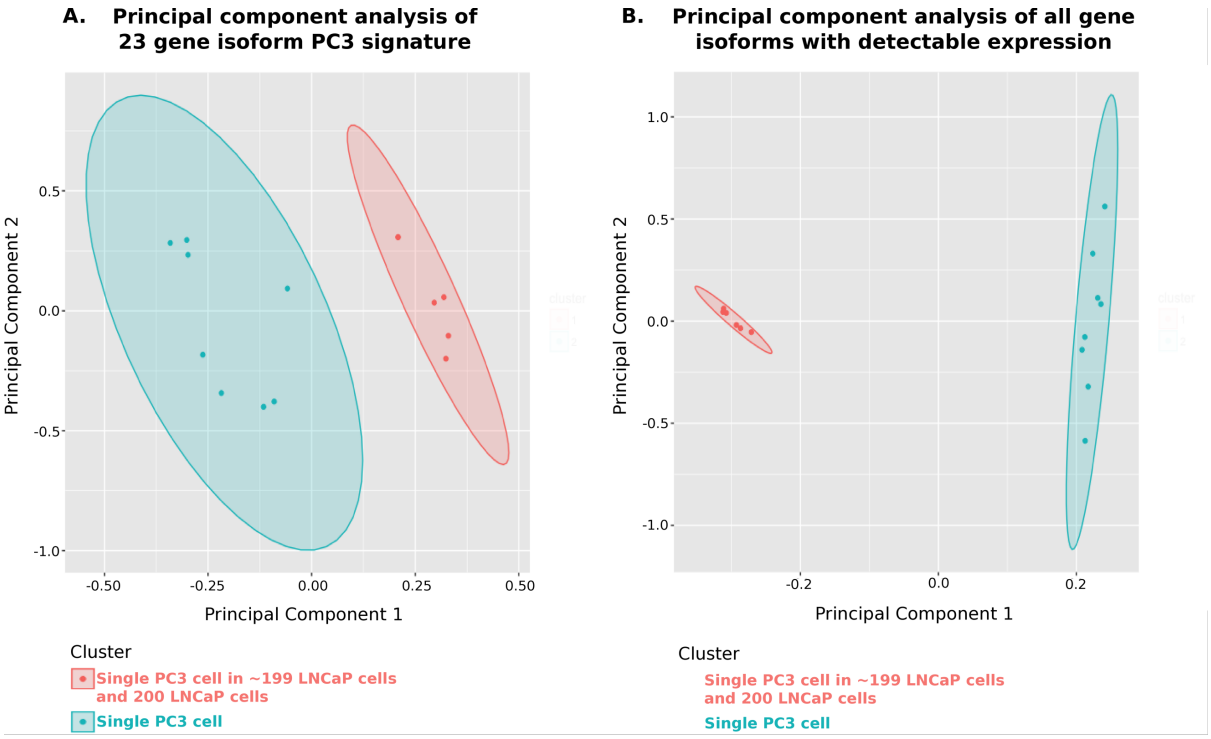

Figure 5.3 – Single PC3 cells cluster independently of the single PC3 cell in 199 LNCaP cells in principal components analysis.

Principal component analysis of samples shows that single PC3 cells in isolation clustered separately from single PC3 cells in 199 LNCaP cells in both the 23 gene isoforms (*k* = 2) and global transcript levels (k = 2) in the PCA/PAM.

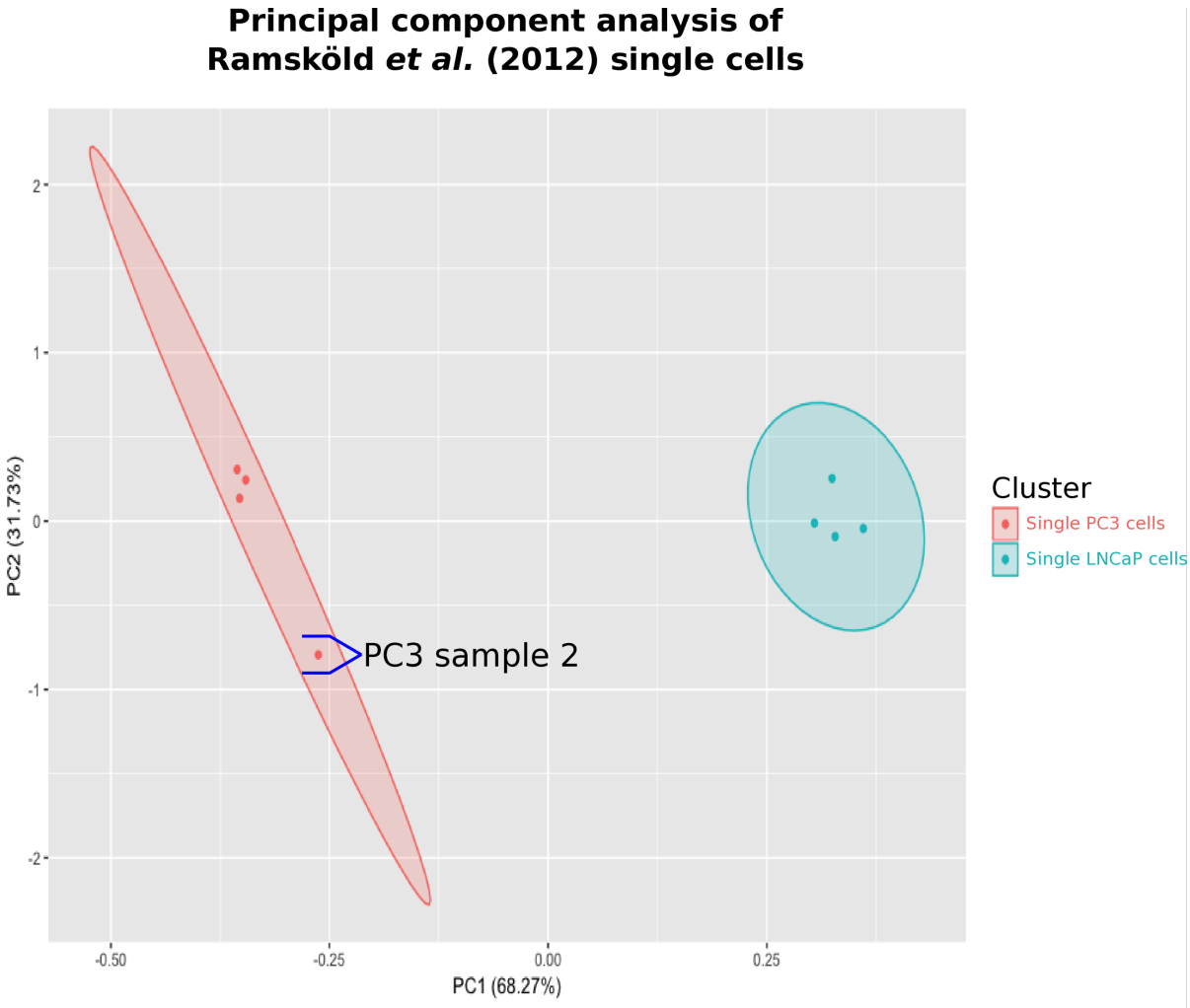

Figure 5.4 – Principal components analysis of Ramsköld *et al.* (2012) single cells.
PCA/PAM clustering of single LNCaP and PC3 cells from Ramsköld *et al.* (2012). Data points are shown based on the first two principal components (PC1 and PC2) that capture the most variance and clustered according to the PAM algorithm (*k* = 2). Percentage values represent the degree to which each principal component accounts for variance. Ellipses assume a multivariate normal distribution with the confidence level set at 95%.

5.3 MA plots

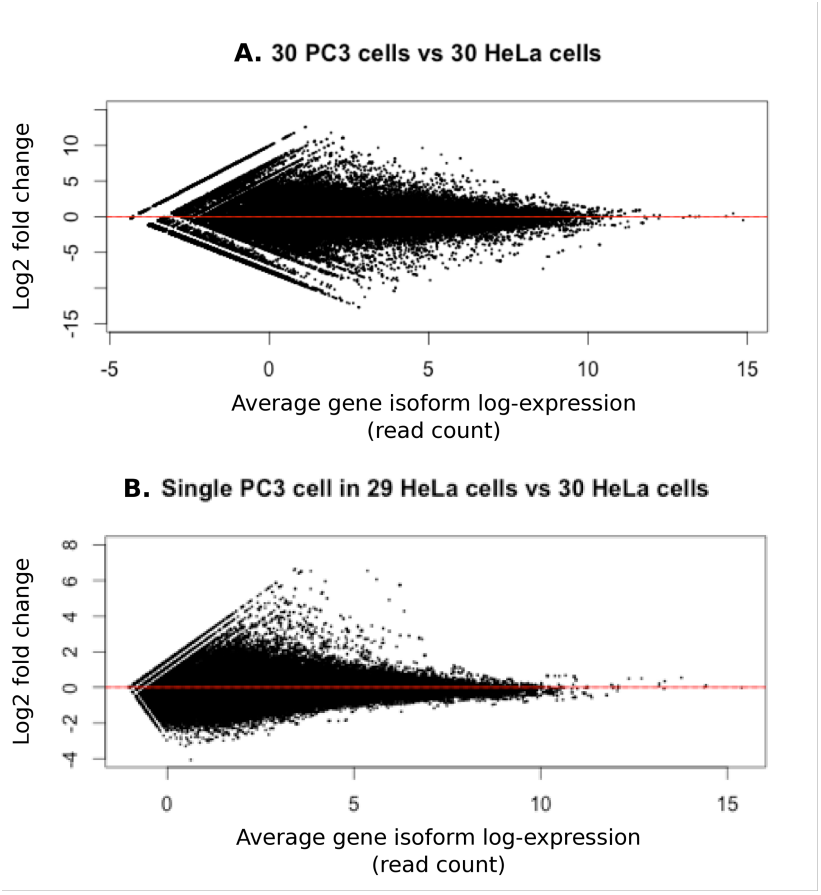

Figure 5.5 – MA plots for PC3 cells vs HeLa cells.

MA plots for 30 PC3 cells vs 30 HeLa cells (A) and single PC3 cell in 29 HeLa cells vs 30 HeLa (B) that correspond to volcano plots A and B respectively in Figure 3.8. Average gene isoform log-expression is over all samples. The Loess smoothed curve (red line) is horizontal at y axis = 0, indicating the absence of systemic bias.

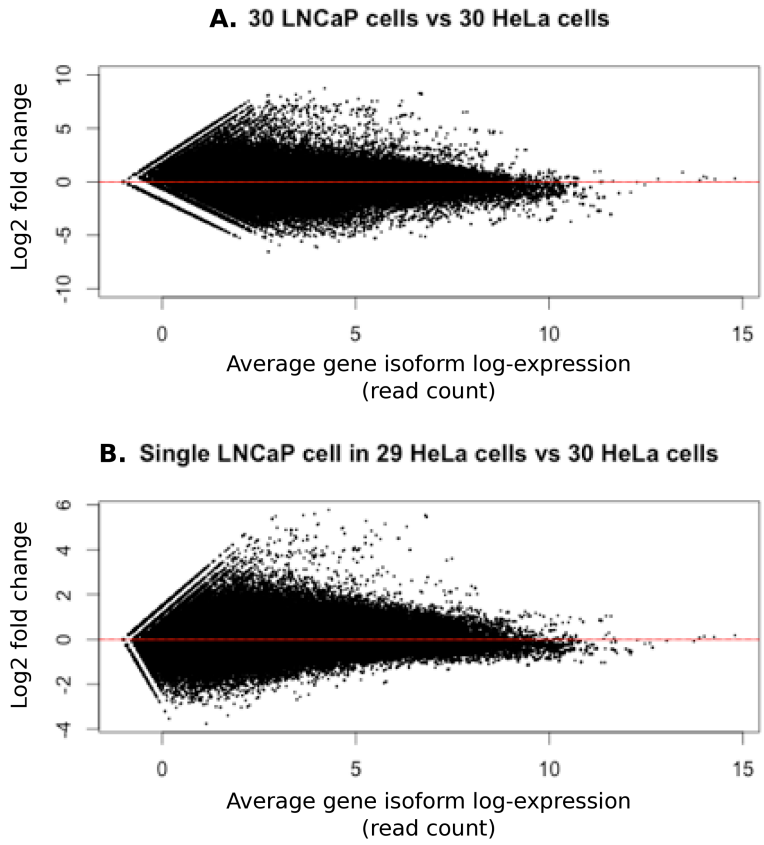

Figure 5.6 – MA plots for LNCaP cells vs HeLa cells.

MA plots for 30 LNCaP cells vs 30 HeLa cells (A) and single LNCaP cell in 29 HeLa cells vs 30 HeLa (B) that correspond to volcano plots A and B respectively in Figure 3.10. Average gene isoform log-expression is over all samples. The Loess smoothed curve (red line) is horizontal at y axis = 0, indicating the absence of systemic bias.

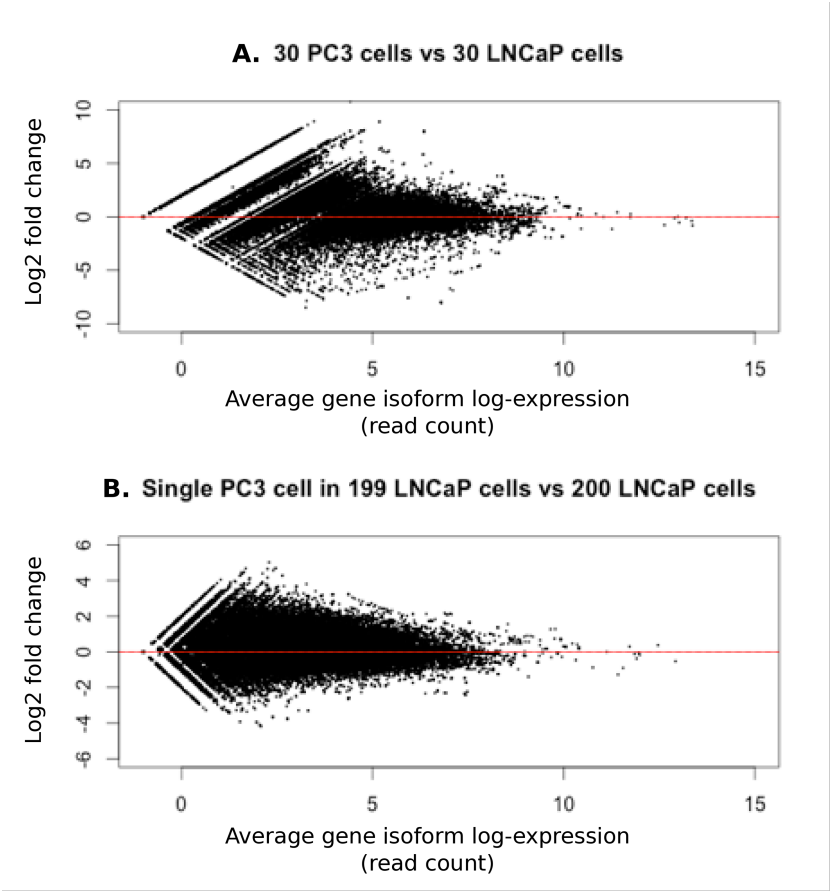

Figure 5.7 – MA plots for LNCaP cells vs PC3 cells.

MA plots for 30 PC3 cells vs 30 LNCaP cells (A) and a single PC3 cell in 199 LNCaP cells vs 200 LNCaP cells (B) that correspond to volcano plots A and B respectively in Figure 3.12. Average gene isoform log-expression is over all samples. The Loess smoothed curve (red line) is horizontal at y axis = 0, indicating the absence of systemic bias.

5.4 mRNA transcript mapping

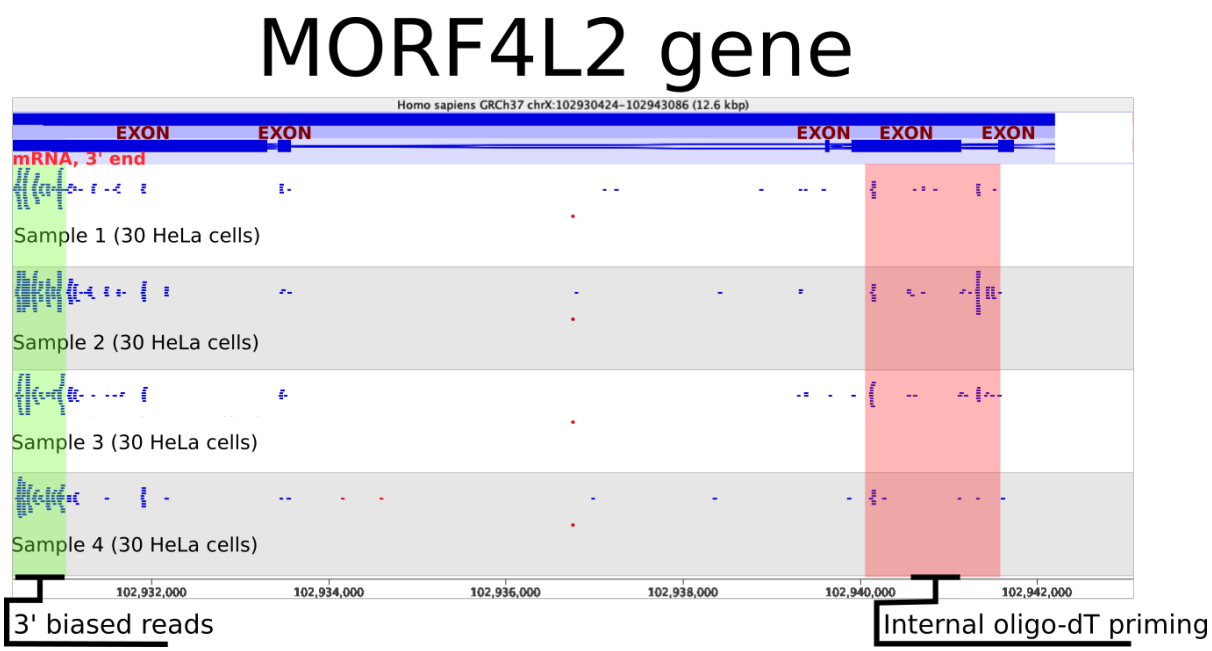

Figure 5.8 – Gene expression view showing internal priming.

Gene expression in MORF4L2 illustrates that RNA-seq reads are predominantly at the 3' end of the mRNA sequence (green region). Reads also align with regions distal from the 3' end, inluding in intronic regions. This is an indication of aberrant internal oligo-dT priming.

5.5 Single cell gene expression detection

Genes detected for single cells

Ramsköld *et al.*, (2012) single cell samples	Genes detected	Total number of genes	Genes detected (%)	Total reads	Genes/reads
LNCaP sample 1	23617	33175	71.18914846	20078740	0.001176219
LNCaP sample 2	23867	33175	71.94272796	22814212	0.001046146
LNCaP sample 3	23556	33175	71.00527506	28759799	0.00081906
LNCaP sample 4	22294	33175	67.20120573	25071173	0.000889228
PC3 sample 1	22885	33175	68.98266767	22711627	0.001007634
PC3 sample 2	22789	33175	68.69329314	22776676	0.001000541
PC3 sample 3	23435	33175	70.64054258	23182050	0.001010911
PC3 sample 4	23861	33175	71.92464205	24029752	0.000992977
	Median = 23495.5		Median = 70.82290882	Median = 22998131	Median = 0.001004088

Table 5.1 – Gene expression in Ramsköld *et al.*, (2012) single cells

Single cell samples	Genes detected	Total number of genes	Genes detected (%)	Total reads	Genes/reads
PC3 sample 1	5501	33175	0.165817634	23462	0.23446424
PC3 sample 2	6127	33175	0.184687265	40235	0.152280353
PC3 sample 3	5639	33175	0.169977393	25012	0.225451783
PC3 sample 4	5618	33175	0.169344386	29074	0.193231066
PC3 sample 5	5657	33175	0.17051997	28605	0.197762629
PC3 sample 6	9167	33175	0.276322532	145815	0.062867332
PC3 sample 7	6394	33175	0.192735494	39390	0.162325463
PC3 sample 8	7301	33175	0.220075358	62305	0.117181607
LNCaP sample 1	5816	33175	0.175312735	36310	0.16017626
LNCaP sample 2	9707	33175	0.292599849	265790	0.036521314
LNCaP sample 3	4853	33175	0.146284853	24102	0.201352585
LNCaP sample 4	5635	33175	0.16985682	34599	0.162865979
LNCaP sample 5	9621	33175	0.290007536	217128	0.044310269
LNCaP sample 6	10816	33175	0.326028636	374678	0.028867454
LNCaP sample 7	10875	33175	0.327807084	258300	0.042102207
LNCaP sample 8	8861	33175	0.267098719	147729	0.059981453
	Median = 6260.5		Median = 0.18871138	Median = 39812.5	Median = 0.156228307

Table 5.2 – Gene expression in single cells under study

Tables 5.1 and 5.2 – Gene expression detection in single cells

Statistics for the number of genes detected with the corresponding number of reads for Ramsköld *et al.*'s (2012) single cells (Table 5.1) and our single cells (Table 5.2).

5.6 Cell cycle genes and single cell genes expressed in a background population

Cell cycle genes and up-regulated genes that distinguish a single cell in background populations

Core 67 cell cycle genes	Single PC3 cell in HeLa	Single LNCaP cell in HeLa	Single PC3 cell in LNCaP
ARHGAP11A	ABI3BP	AC006157.2	ABCD1
BIRC5	AC073995.2	AC009475.2	AEN
BRD8	AGPAT4	AR	ALKBH8
BUB1	AGXT2L2	BEX4	ARHGAP32
BUB1B	ALDH1A2	C20orf108	CLDN20
BUB3	ANXA11	C8orf42	COCH
CCNA2	ARHGAP29	CTD-2314B22.3	ENTPD7
CCNB1	BCMO1	DDC	EPB41L2
CCNB2	C10orf55	DMXL1	FBXL13
CCNE1	CERS6	FOLH1	HSPD1
CCNE2	CHRM3	HOMER2	LAPTM4B
CCNF	CLDN14	KLK3 (PSA)	PTPN21
CDC20	CMTM7	LCP1	RP11-829H16.3
CDC25B	CNN3	MAOA	SYNJ2
CDC25C	COL13A1	MIPOL1	TNIK
CDC6	COL6A1	ODZ1	U52111.14
CDCA3	CSMD2	PRAC	VPS29
CDCA8	DPP4	RPS4Y1	WFDC3
CENPE	DSC3	STEAP2	
CENPF	DUSP6	TMPRSS2	
CHAF1A	F2RL1	UTRN	
CKAP2	FAM43A	ZBTB10	
CKS1B	FAM84B	ZIM2	
CKS2	FER1L6		
DTL	FOSL1		
E2F1	FOXP1		
ESPL1	G0S2		
EXO1	GNG2		
FAM64A	GPR110		
GPR126	HAS3		
GPSM2	HIPK2		
GTSE1	HMGA2		
H2AFX	HOXB13		
HMGB2	HTR1E		

Core 67	PC3 in HeLa
HMGB3	KIF5C
HMMR	KRT19
HSPB8	LPAR1
KIF11	MAGED1
KIF22	ME1
KIF23	NDST3
KIF2C	NEFL
KPNA2	NES
LBR	NFIX
MCM2	NRP1
MCM6	NTNG1
MKI67	OAT
NDE1	PHLDA1
NEK2	PKP1
NUSAP1	PLAU
PASK	PRAC
PCNA	PRSS3
PLK1	PSD3
POLD3	RNF144A
PRC1	RP11-117P22.1
PTTG1	RP11-133O22.6
RFC4	RP11-463J7.2
SFPQ	RP11-65F13.2
SLBP	RP11-66B24.4
SPAG5	SATB1
TACC3	SCOC
TOP2A	SHISA3
TPX2	SPATS2L
TROAP	SPINT2
TTK	SULF2
UBE2C	SYT1
UBE2S	TMEM45B
UNG	TRIM58
	TSPAN5
	UCHL1
	ZNF503

Table 5.3 – Genes over-expressed in single cells relative to a background population

List of genes from the 'core 67' cell cycle genes (Dominguez *et al*., 2016) alongside genes that comprise the gene isoforms which distinguish single PC3 and LNCaP cells in a HeLa or LNCaP background (Figures 3.8, 3.10 and 3.12, $P < 0.05$).

5.7 Cumulative distribution plot for single cells

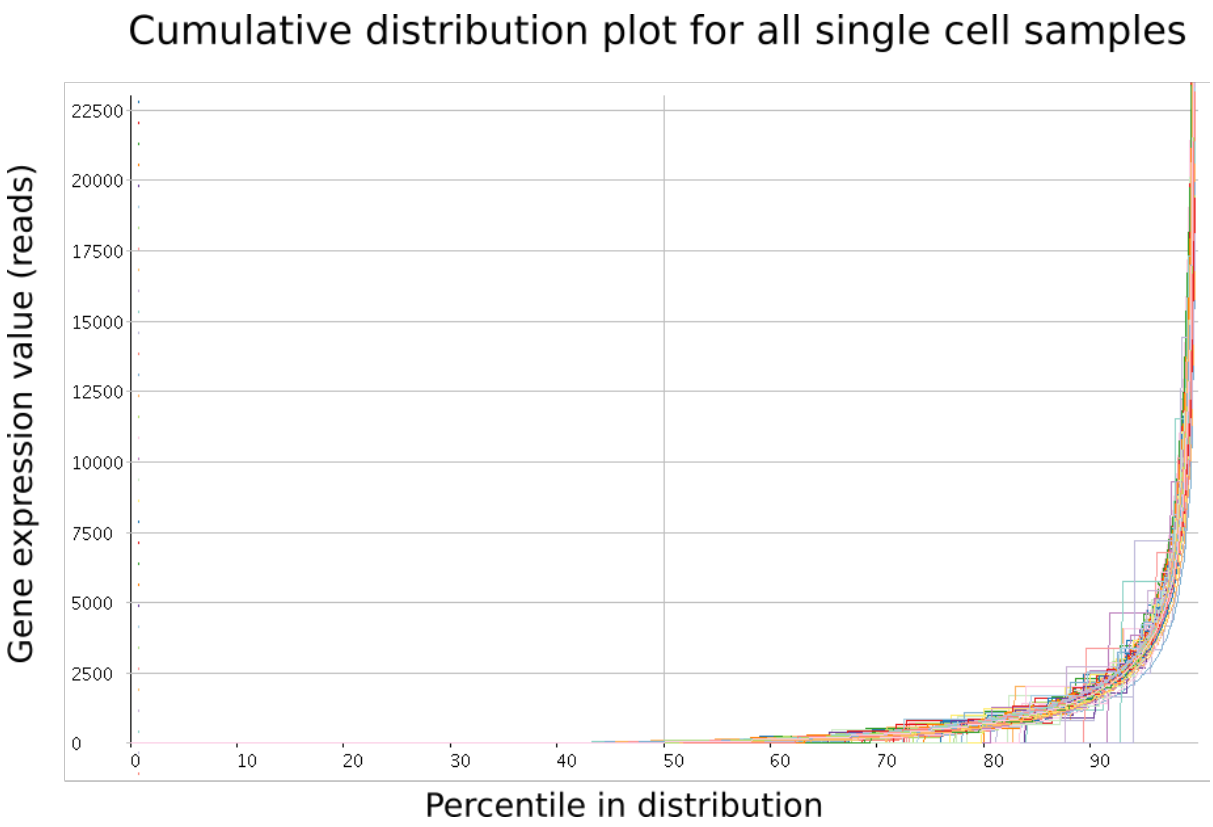

Figure 5.9 – Cumulative distribution plot for all single cells.

This plot depicts the distribution profiles of all gene expression values across all single cells (Ramsköld *et al*. and ours) after scaling the read counts of all samples to the sample with the highest overall read count. The distribution profiles of all samples are clutered together, demonstrating that the samples are well normalised.

5.8 Dendrogram for single PC3 cell in 199 LNCaP cells experiment

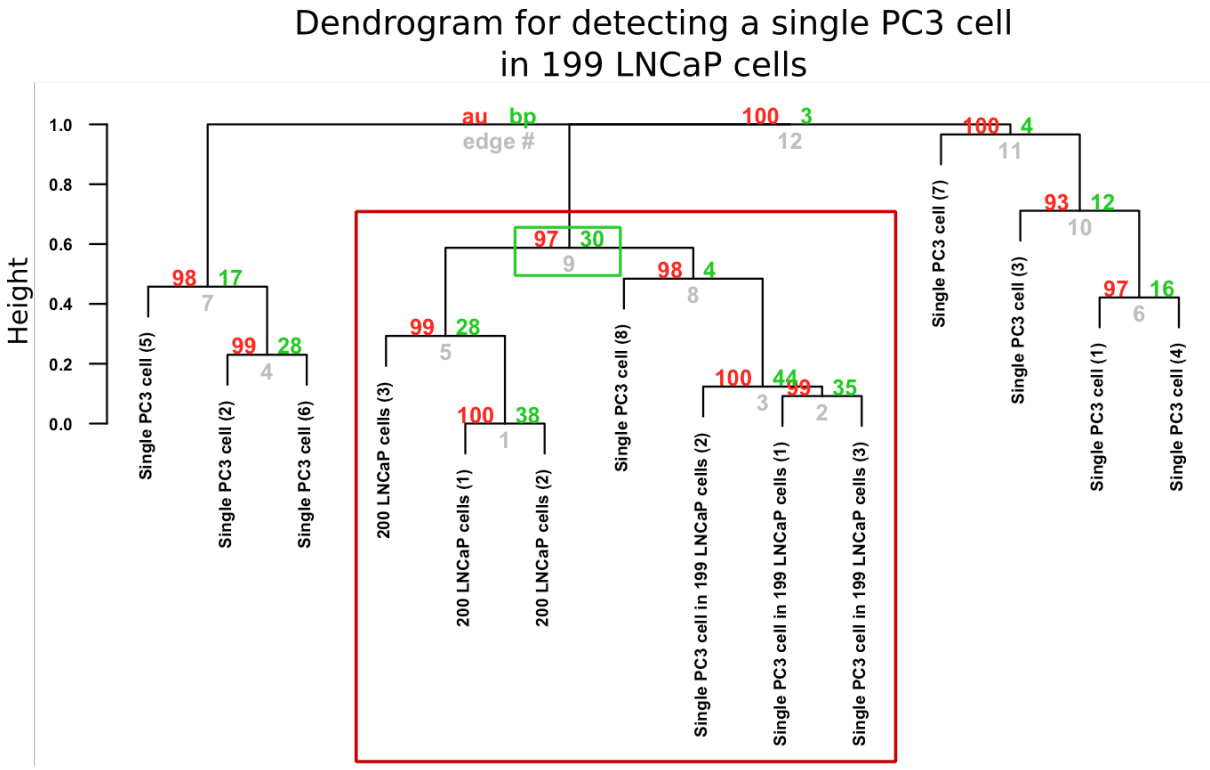

Figure 5.10 – Dendrogram for detecting a single PC3 cell in 199 LNCaP cells.

This plot depicts the statistical separation between 200 LNCaP cells, a single PC3 cell in 199 LNCaP cells and single PC3 cells alone. Node numbers ('edge #', in gray), bootstrap values ('bp', in green) and corresponding approximately unbiased P values ('au', in red) are shown for each node. The complete linkage method was used for hierarchical clustering. The uncentered correlation function was used for the distance measure. Bootstrap values were generated from 10,000 bootstrap replications.

5.9 Reagent preparation

5.8.1 Cryopreservation

The freezing medium for cryopreservation of all cell cultures consisted of 95% complete growth medium with 5% DMSO.

5.8.2 Trypsin preparation

0.5% trypsin was diluted in PBS 1:10 for 0.05% trypsin.

5.8.2 PBS preparation

PBS was prepared by adding 1 PBS tablet per 100 mL in mQ H_2O, then autoclaved for sterilization

5.10 R packages and code

5.10.1 Heatmap code

#Read in data

>library(gplots)

>data<-read.csv('gene_expression_file_from_seq_monk.csv')

#Remove genes without expression across all samples in file

>data_mat<-data[apply(data[,2:ncol(data)], 1, sum)>0,]

#Normalise data by the standard deviation

>data_norm<-t(apply(data_mat2, 1, function(x) (x-mean(x))/sd(x)))

#Plot heatmap

>heatmap.2(data_norm, distfun = function(x) as.dist(1-cor(t(x), method = c("spearman"))),
Rowv = T, dendrogram='column', col =colorRampPalette(c("green","white","red"))(100),
scale='row', trace='none', margins=c(2,4), ylab='Genes', cex.lab=0.75, cexCol=0.8, srtCol=20,
lwid=c(1,3.5))

5.10.2 Volcano plot code

#Read in data

>library(limma)

>data<-read.csv("gene_expression_file_from_seq_monk.csv")

#Select only gene expression data from file generated by SeqMonk

>x<-data[,grep("Mean", names(dat))]

#Convert data to log2 scale

>x[x==0]<-0.5

>x<-log2(x)

#Create data frame (e.g. 6 samples total, 3 vs 3)

>dd<-cbind(rep(1,6),c(1,1,1,0,0,0))

>fit<-eBayes(lmFit(x,dd))

>tt<-topTable(fit,n=149135,coef=2)

#Produce volcano plot

>with(tt, plot(logFC, -log(adj.P.Val),pch=9,cex=0.3,col="black", xlab="Log2 Fold Change", ylab="-log Adjusted P value", main="title", ylim=c(0,8), xlim=c(-6,6)))

>with(subset(tt, adj.P.Val<.05 & logFC>1), points(logFC, -log(adj.P.Val), pch=9,cex=0.3, col="red"))

>with(subset(tt, adj.P.Val<.05 & -logFC>1), points(logFC, -log(adj.P.Val), pch=9,cex=0.3, col="green"))

>abline(h=-log(0.05),v=1,col='blue',lty=2)

>abline(v=-1,col='blue',lty=2)

#Produce MA plot

>plotMA(fit, main='title', ylab='Log2 Fold Change', xlab='Average Expression')

>abline(h=0,col='red',lty=1)

5.10.3 Principal component analysis code

#Read in data

>library(gplots)

>library(ggfortify)

>library(cluster)

>data<-read.csv('gene_expression_file_from_seq_monk.csv')

#Remove genes with zero expression across all samples in file

>data_mat<-data[apply(data[,2:ncol(data)], 1, sum)>0,]

#Convert data to log2 scale and normalize by the standard deviation

>pca = apply(data_mat,2,function(x){

x[x==0] = 1

x

l = log2(x)

(l - mean(l))/sd(l)

})

#Generate first two principal components

>pc = prcomp(t(pca))

>scores <- data.frame(pc$x[,1:2])

#Clustering by partitioning around the medoids

*>autoplot(pam(scores[-3],*input k-value, e.g. 2*), frame = FALSE, frame.type = 'norm')*

5.10.4 Dendrogram and bootstrap analysis code

>library(pvclust)

> file<-read.csv('gene_expression_file.csv')

>result <- pvclust(file, method.dist="uncentered", method.hclust="complete", nboot=10000, parallel=TRUE)

>plot(result)

Acknowledgements

I would like to thank my supervisor, Professor Parry Guilford for his guidance, patience and kind support. Every meeting with Professor Guilford left me inspired to confront the next challenge. His editing of my work was extremely appreciated.

I am particularly indebted to my advisor, Dr. Robert Day, who was kind enough to take me under his wing in the lab and show me how hard science is done. His mentorship through my methods and results chapters was extremely appreciated. I am also thankful for his encouragement and support for submitting a poster at the Australasian Genomic Technologies Association conference 2016. I cannot thank him enough.

I am thankful for the assistance and wealth of technical experience of Tanis Godwin, Dr. Augustine Chen and Dr. Donghui Zhou. Special thanks to Tanis Godwin for mentoring me in cell culture and her effort and skill in isolating the single cells. I am grateful to all the members in the Cancer Genetics Laboratory who created an atmosphere of support and critical thinking.

I am grateful to Associate Professor Mik Black, who supervised my fourth year project and introduced me to Professor Guilford. He was always ready and available to provide me with gems of statistical and biological knowledge.

I am also grateful to Dr. Elspeth Gold who provided me with the LNCaP cell line. She sadly passed away in 2015, a great loss for prostate cancer research. I will remember her extreme readiness to help in any capacity.

I would like to thank my parents for all their support, without which this thesis would not have been possible. I thank my mother for all her support through her trials, she is the most amazing person I know.

List of References

Andriole, G. L., Levin, D. L., Crawford, E. D., Gelmann, E. P., Pinsky, P. F., Chia, D., … for the PLCO, P. T. (2005). Prostate Cancer Screening in the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial: Findings From the Initial Screening Round of a Randomized Trial. *JNCI Cancer Spectrum*, *97*(6), 433–438.

Bahar Halpern, K., Tanami, S., Landen, S., Chapal, M., Szlak, L., Hutzler, A., … Itzkovitz, S. (2015). Bursty gene expression in the intact mammalian liver. *Molecular Cell*, *58*(1), 147–156. https://doi.org/10.1016/j.molcel.2015.01.027

Baran-Gale, J., Chandra, T., & Kirschner, K. (2018). Experimental design for single-cell RNA sequencing. *Briefings in Functional Genomics*, *17*(4), 233–239. https://doi.org/10.1093/bfgp/elx035

Barron, M., & Li, J. (2016). Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data. *Scientific Reports*, *6*. https://doi.org/10.1038/srep33892

Basch, E., Oliver, T. K., Vickers, A., Thompson, I., Kantoff, P., Parnes, H., … Nam, R. K. (2012). Screening for Prostate Cancer with Prostate-Specific Antigen Testing: American Society of Clinical Oncology Provisional Clinical Opinion. *Journal of Clinical Oncology*. https://doi.org/10.1200/JCO.2012.43.3441

Benjamini Y. and Hochberg Y. (1995). Controlling the FDR:A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B*, *57*(1), 289–300.

Berney, D. (2010). Biomarkers for prostate cancer detection and progression: Beyond prostate-specific antigen. *Drug News & Perspectives*, *23*(3), 185. https://doi.org/10.1358/dnp.2010.23.3.1437708

Bill-Axelson, A., Holmberg, L., Garmo, H., Rider, J. R., Taari, K., Busch, C., … Johansson, J. E. (2014). Radical prostatectomy or watchful waiting in early prostate cancer. *N Engl J Med*, *370*(10), 932–942. https://doi.org/10.1056/NEJMoa1311593

Bolon, B., & Graham, D. G. (2011). Fundamental Neuropathology for Pathologists and Toxicologists: An Introduction. In *Fundamental Neuropathology for Pathologists and Toxicologists* (pp. 1–14). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470939956.ch1

Brady, G., Barbara, M., & Iscove, N. N. (1990). *Representative in Vitro cDNA Amplification From Individual Hemopoietic Cells and Colonies*. *METHODS IN MOLECULAR AND CELLULAR BIOLOGY* (Vol. 2). Retrieved from https://pdfs.semanticscholar.org/e44b/76e215ed1cb865bbc72295b425107ad72273.pdf

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., … Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, *10*(11), 1093–1098. https://doi.org/10.1038/nmeth.2645

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., … Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, *10*(11), 1093–1098. https://doi.org/10.1038/nmeth.2645

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., … Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, *33*(2), 155–160. https://doi.org/10.1038/nbt.3102

Catalona, W. J., Richie, J. P., Ahmann, F. R., Hudson, M. A., Scardino, P. T., Flanigan, R. C., … Southwick, P. C. (2017). Comparison of Digital Rectal Examination and Serum Prostate Specific Antigen in the Early Detection of Prostate Cancer: Results of a Multicenter Clinical Trial of 6,630 Men. *Journal of Urology*, *197*(2), S200–S207. https://doi.org/10.1016/j.juro.2016.10.073

Carter, H. B., Albertsen, P. C., Barry, M. J., Etzioni, R., Freedland, S. J., Greene, K. L., … Zietman, A. L. (2013). Early detection of prostate cancer: AUA guideline. *Journal of Urology*, *190*(2), 419–426. https://doi.org/10.1016/j.juro.2013.04.119

Chapman, A. R., He, Z., Lu, S., Yong, J., Tan, L., Tang, F., & Xie, X. S. (2015). Single cell transcriptome amplification with MALBAC. *PLoS ONE*, *10*(3). https://doi.org/10.1371/journal.pone.0120889

Choi, Y. H., & Kim, J. K. (2019). Dissecting cellular heterogeneity using single-cell RNA sequencing. *Molecules and Cells*. https://doi.org/10.14348/molcells.2019.2446

Chu, T. M., Murphy, G. P., Kawinski, E., & Mirand, E. A. (1983). LNCaP model of human prostatic carcinoma. *Cancer Research*, *43*(4), 1809–1818.

Clare, A. J., Day, R. C., Empson, R. M., & Hughes, S. M. (2018). Transcriptome profiling of layer 5 intratelencephalic projection neurons from the mature mouse motor cortex. *Frontiers in Molecular Neuroscience*, *11*. https://doi.org/10.3389/fnmol.2018.00410

Cleveland, W. S., Grosse, E., & Shyu, W. M. (2017). Local regression models. In *Statistical Models in S* (pp. 309–376). https://doi.org/10.1201/9780203738535

Cohen, R. J., Shannon, B. A., Phillips, M., Moorin, R. E., Wheeler, T. M., & Garrett, K. L. (2008). Central Zone Carcinoma of the Prostate Gland: A Distinct Tumor Type With Poor Prognostic Features. *Journal of Urology*, *179*(5), 1762–1767. https://doi.org/10.1016/j.juro.2008.01.017

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*. https://doi.org/10.1186/s13059-016-0881-8

Craft, N., Chhor, C., Tran, C., Belldegrun, A., DeKernion, J., Witte, O. N., … Sawyers, C. L. (1999). Evidence for clonal outgrowth of androgen-independent prostate cancer cells from androgen-dependent tumors through a two-step process. *Cancer Research*, *59*(19), 5030–5036.

Dall'Era, M. A., Maddala, T., Polychronopoulos, L., Gallagher, J. R., Febbo, P. G., & Denes, B. S. (2015). Utility of the Oncotype DX® Prostate Cancer Assay in Clinical Practice for Treatment Selection in Men Newly Diagnosed with Prostate Cancer: A Retrospective Chart Review Analysis. *Urology Practice*, *2*(6), 343–348. https://doi.org/10.1016/j.urpr.2015.02.007

Day, R. C., Godwin, T. D., Harris, C., Stockwell, P., Shaw, A., & Guilford, P. J. (2018). Integrated Cell Expression Toolkit for low input transcript profiling. *bioRxiv*, 458851. https://doi.org/10.1101/458851

Day, R. C., McNoe, L., & Macknight, R. C. (2007). Evaluation of global RNA amplification and its use for high-throughput transcript analysis of laser-microdissected endosperm. *International Journal of Plant Genomics*, *2007*. https://doi.org/10.1155/2007/61028

Djavan, B., Mazal, P., Zlotta, A., Wammack, R., Ravery, V., Remzi, M., … Marberger, M. (2001). Pathological features of prostate cancer detected on initial and repeat prostate biopsy: Results of the prospective European prostate cancer detection study. *Prostate*, *47*(2), 111–117. https://doi.org/10.1002/pros.1053

Dominguez, D., Tsai, Y. H., Gomez, N., Jha, D. K., Davis, I., & Wang, Z. (2016). A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and cancer. *Cell Research*, *26*(8), 946–962. https://doi.org/10.1038/cr.2016.84

Duijvesz, D., Luider, T., Bangma, C. H., & Jenster, G. (2011). Exosomes as biomarker treasure chests for prostate cancer. *European Urology*, *59*(5), 823–831. https://doi.org/10.1016/j.eururo.2010.12.031

Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., … Coleman, P. (1992). Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(7), 3010–3014. https://doi.org/10.1073/pnas.89.7.3010

Ferlay, J., Parkin, D. M., & Steliarova-Foucher, E. (2010). Estimates of cancer incidence and mortality in Europe in 2008. *European Journal of Cancer*, *46*(4), 765–781. https://doi.org/10.1016/j.ejca.2009.12.014

Ferreira, L. B., Palumbo, A., de Mello, K. D., Sternberg, C., Caetano, M. S., de Oliveira, F. L., … Gimba, E. R. P. (2012). PCA3 noncoding RNA is involved in the control of prostate-cancer cell survival and modulates androgen receptor signaling. *BMC Cancer*, *12*. https://doi.org/10.1186/1471-2407-12-507

Fischer, O. M., Streit, S., Hart, S., & Ullrich, A. (2003). Beyond Herceptin and Gleevec. *Current Opinion in Chemical Biology*. https://doi.org/10.1016/S1367-5931(03)00082-6

Foot, N. C., Papanicolaou, G. N., Holmquist, N. D., & Seybolt, J. F. (1958). Exfoliative cytology of urinary sediments. A review of 2,829 cases. *Cancer*, *11*(1), 127–137. https://doi.org/10.1002/1097-0142(195801/02)11:1<127::AID-CNCR2820110124>3.0.CO;2-W

Frisch, S. M., Schupp, J., Killiam, E., Kumar, S., Rimm, D. L., Cieply, B., … Park, S. H. (2011). A Pathway for the Control of Anoikis Sensitivity by E-Cadherin and Epithelial-to-Mesenchymal Transition. *Molecular and Cellular Biology*, *31*(19), 4036–4051. https://doi.org/10.1128/mcb.01342-10

Fujita, K., Pavlovich, C. P., Netto, G. J., Konishi, Y., Isaacs, W. B., Ali, S., … Meeker, A. K. (2009). Specific detection of prostate cancer cells in urine by multiplex immunofluorescence cytology. *Human Pathology*, *40*(7), 924–933. https://doi.org/10.1016/j.humpath.2009.01.004

Griffiths, J. A., Scialdone, A., & Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular Systems Biology*, *14*(4). https://doi.org/10.15252/msb.20178046

Gross, A., Schoendube, J., Zimmermann, S., Steeb, M., Zengerle, R., & Koltay, P. (2015). Technologies for single-cell isolation. *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms160816897

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., … Yanai, I. (2016). CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0938-8

Hessels, D., Klein Gunnewiek, J. M. T., Van Oort, I., Karthaus, H. F. M., Van Leenders, G. J. L., Van Balken, B., … Culig, Z. (2003). DD3PCA3-based molecular urine analysis for the diagnosis of prostate cancer. *European Urology*, *44*(1), 8–16. https://doi.org/10.1016/S0302-2838(03)00201-X

Hume, D. A. (2008). Differentiation and heterogeneity in the mononuclear phagocyte system. *Mucosal Immunology*. https://doi.org/10.1038/mi.2008.36

Iscove, N. N., Barbara, M., Gu, M., Gibson, M., Modi, C., & Winegarden, N. (2002). Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nature Biotechnology*, *20*(9), 940–943. https://doi.org/10.1038/nbt729

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., … Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, *11*(2), 163–166. https://doi.org/10.1038/nmeth.2772

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J. B., Lönnerberg, P., & Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, *21*(7), 1160–1167. https://doi.org/10.1101/gr.110882.110

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., … Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, *343*(6172), 776–779. https://doi.org/10.1126/science.1247651

Jeffreys, A. J., Wilson, V., Neumann, R., & Keyte, J. (1988). Amplification of human minisatellites by the polymerase chain reaction: Towards DNA fingerprinting of single cells. *Nucleic Acids Research*, *16*(23), 10953–10971. https://doi.org/10.1093/nar/16.23.10953

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: A Cancer Journal for Clinicians*, *61*(2), 69–90. https://doi.org/10.3322/caac.20107

Jemal, A., Center, M. M., DeSantis, C., & Ward, E. M. (2010). Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiology Biomarkers and Prevention*. https://doi.org/10.1158/1055-9965.EPI-10-0437

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., … Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, *21*(9), 1543–1551. https://doi.org/10.1101/gr.121095.111

Kaighn M., Narayan K. S., Ohnuki Y., Lechner J., Jones L., (1979). Establishment and characterization of a human prostatic carcinoma cell line (PC-3). Invest. Urol. 17: 16–23.

Karantanos, T., Corn, P. G., & Thompson, T. C. (2013). Prostate cancer progression after androgen deprivation therapy: Mechanisms of castrate resistance and novel therapeutic approaches. *Oncogene*. https://doi.org/10.1038/onc.2013.206

Kaufmann, B. B., & van Oudenaarden, A. (2007). Stochastic gene expression: from single molecules to the proteome. *Current Opinion in Genetics and Development*. https://doi.org/10.1016/j.gde.2007.02.007

Keetch, D. W., Catalona, W. J., & Smith, D. S. (1994). Serial prostatic biopsies in men with persistently elevated serum prostate specific antigen values. *Journal of Urology*, *151*(6), 1571–1574. https://doi.org/10.1016/S0022-5347(17)35304-1

Kumar, S., Park, S. H., Cieply, B., Schupp, J., Killiam, E., Rimm, D. L., & Frisch, S. M. (2011). A pathway for the control of anoikis-sensitivity by E-cadherin and EMT. *Mol. Cell. Biol.*, MCB.01342-10. https://doi.org/10.1128/mcb.01342-10

Kurimoto, K., Yabuta, Y., Ohinata, Y., Ono, Y., Uno, K. D., Yamada, R. G., … Saitou, M. (2006). An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Research*, *34*(5), e42. https://doi.org/10.1093/nar/gkl050

Kwist, K., Bridges, W. C., & Burg, K. J. L. (2016). The effect of cell passage number on osteogenic and adipogenic characteristics of D1 cells. *Cytotechnology*, *68*(4), 1661–1667. https://doi.org/10.1007/s10616-015-9883-8

Lang, R., Liu, G., Shi, Y., Bharadwaj, S., Leng, X., Zhou, X., … Zhang, Y. (2013). Self-Renewal and Differentiation Capacity of Urine-Derived Stem Cells after Urine Preservation for 24 Hours. *PLoS ONE*, *8*(1). https://doi.org/10.1371/journal.pone.0053980

Landry, J. J. M., Pyl, P. T., Rausch, T., Zichner, T., Tekkedil, M. M., Stütz, A. M., … Steinmetz, L. M. (2013). The Genomic and Transcriptomic Landscape of a HeLa Cell Line. *G3: Genes|Genomes|Genetics*, *3*(8), 1213–1224. https://doi.org/10.1534/g3.113.005777

Landry, J. J. M., Pyl, P. T., Rausch, T., Zichner, T., Tekkedil, M. M., Stütz, A. M., … Steinmetz, L. M. (2013). The Genomic and Transcriptomic Landscape of a HeLa Cell Line. *G3: Genes|Genomes|Genetics*, *3*(8), 1213–1224. https://doi.org/10.1534/g3.113.005777

Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N., & Werb, Z. (2018). Tumour heterogeneity and metastasis at single-cell resolution. *Nature Cell Biology*. https://doi.org/10.1038/s41556-018-0236-7

Lee, D., Cheng, A., Lawlor, N., Bolisetty, M., & Ucar, D. (2018). Detection of correlated hidden factors from single cell transcriptomes using Iteratively Adjusted-SVA (IA-SVA). *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-35365-9

Lee, J. J., Thomas, I. C., Nolley, R., Ferrari, M., Brooks, J. D., & Leppert, J. T. (2015). Biologic differences between peripheral and transition zone prostate cancer. *Prostate*, *75*(2), 183–190. https://doi.org/10.1002/pros.22903

Leyten, G. H. J. M., Hessels, D., Jannink, S. A., Smit, F. P., De Jong, H., Cornel, E. B., … Schalken, J. A. (2014). Prospective multicentre evaluation of PCA3 and TMPRSS2-ERG gene fusions as diagnostic and prognostic urinary biomarkers for prostate cancer. *European Urology*, *65*(3), 534–542. https://doi.org/10.1016/j.eururo.2012.11.014

Li, B., Hartono, C., Ding, R., Sharma, V. K., Ramaswamy, R., Qian, B., … Suthanthiran, M. (2002). Noninvasive Diagnosis of Renal-Allograft Rejection by Measurement of Messenger RNA for Perforin and Granzyme B in Urine. *New England Journal of Medicine*, *344*(13), 947–954. https://doi.org/10.1056/nejm200103293441301

Liu, J., Wang, C.-M., Chen, C.-L., Huang, S., Huang, T., & Lin, C.-L. (2015). Abstract 4569: Digitizing single-cell expression patterns in urine for prostate cancer detection. *Cancer Research*, *74*(19 Supplement), 4569–4569. https://doi.org/10.1158/1538-7445.am2014-4569

Loeb, S., Carter, H. B., Catalona, W. J., Moul, J. W., & Schroder, F. H. (2012). Baseline prostate-specific antigen testing at a young age. *European Urology*. https://doi.org/10.1016/j.eururo.2011.07.067

Luo, T., Fan, L., Zhu, R., & Sun, D. (2019). Microfluidic single-cell manipulation and analysis: Methods and applications. *Micromachines*. https://doi.org/10.3390/mi10020104

Lun, A. T. L., Bach, K., & Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0947-7

Marinov, G. K., Williams, B. A., McCue, K., Schroth, G. P., Gertz, J., Myers, R. M., & Wold, B. J. (2014). From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Research*, *24*(3), 496–510. https://doi.org/10.1101/gr.161034.113

Matz, M., Shagin, D., Bogdanova, E., Britanova, O., Lukyanov, S., Diatchenko, L., & Chenchik, A. (1999). Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Research*, *27*(6), 1558–1560. https://doi.org/10.1093/nar/27.6.1558

McDavid, A., Finak, G., & Gottardo, R. (2016). The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. *Nature Biotechnology*, *34*(6), 591–593. https://doi.org/10.1038/nbt.3498

McNeal, J. E. (1988). Normal anatomy of the prostate and changes in benign prostatic hypertrophy and carcinoma. *Seminars in Ultrasound, CT, and MR*, *9*(5), 329–334. Retrieved from http://europepmc.org/abstract/MED/2483527

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, *5*(10). https://doi.org/10.1371/journal.pgen.1000686

Melamud, E., & Moult, J. (2009). Stochastic noise in splicing machinery. *Nucleic Acids Research*, *37*(14), 4873–4886. https://doi.org/10.1093/nar/gkp471

Ministry of Health (2011). *Cancer: New registrations and deaths 2008*. Wellington: Ministry of Health.

Moreton, J., Dunham, S. P., & Emes, R. D. (2014). A consensus approach to vertebrate de novo transcriptome assembly from RNA-seq data: assembly of the duck (Anas platyrhynchos) transcriptome . *Frontiers in Genetics* . Retrieved from https://www.frontiersin.org/article/10.3389/fgene.2014.00190

Morris, J., Singh, J. M., & Eberwine, J. H. (2011). Transcriptome Analysis of Single Cells. *Journal of Visualized Experiments*, (50). https://doi.org/10.3791/2634

Mostafavi, S., Battle, A., Zhu, X., Urban, A. E., Levinson, D., Montgomery, S. B., & Koller, D. (2013). Normalizing RNA-Sequencing Data by Modeling Hidden Covariates with Prior Knowledge. *PLoS ONE*, *8*(7). https://doi.org/10.1371/journal.pone.0068141

Moyed, H. S., & Bertrand, K. P. (1983). hipA, a newly recognized gene of Escherichia coli K-12 that affects frequency of persistence after inhibition of murein synthesis. *Journal of Bacteriology*, *155*(2), 768–75. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6348026%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC217749

Nam, D. K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., … Wang, S. M. (2002). Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proceedings of the National Academy of Sciences*, *99*(9), 6152–6156. https://doi.org/10.1073/pnas.092140899

Nazir, B. (2014). Pain during transrectal ultrasound-guided prostate biopsy and the role of periprostatic nerve block: What radiologists should know. *Korean Journal of Radiology*. https://doi.org/10.3348/kjr.2014.15.5.543

Neveu, B., Jain, P., Têtu, B., Wu, L., Fradet, Y., & Pouliot, F. (2015). A *PCA3* gene-based transcriptional amplification system targeting primary prostate cancer. *Oncotarget*, *7*(2). https://doi.org/10.18632/oncotarget.6360

Nickens, K. P., Ali, A., Scoggin, T., Tan, S. H., Ravindranath, L., McLeod, D. G., … Petrovics, G. (2015). Prostate cancer marker panel with single cell sensitivity in urine. *Prostate*, *75*(9), 969–975. https://doi.org/10.1002/pros.22981

Ohnuki, Y., Marnell, M. M., Babcock, M. S., Lechner, J. F., & Kaighn, M. E. (1980). Chromosomal Analysis of Human Prostatic Adenocarcinoma Cell Lines. *Cancer Research*, *40*(3), 524–534.

O'Flanagan, C. H., Campbell, K. R., Zhang, A. W., Lim, J. L. P., Biele, J., Eirew, P., … Andy, J. (2019). Dissociation of solid tumour tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Bioarxiv*, 683227. https://doi.org/10.1101/683227

Padovan-Merhar, O., Nair, G. P., Biaesch, A. G., Mayer, A., Scarfone, S., Foley, S. W., … Raj, A. (2015). Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms. *Molecular Cell*, *58*(2), 339–352. https://doi.org/10.1016/j.molcel.2015.03.005

Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., & Smyth, G. K. (2016). Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Annals of Applied Statistics*, *10*(2), 946–963. https://doi.org/10.1214/16-AOAS920

Picelli, S. (2017). Single-cell RNA-sequencing: The future of genome biology is now. *RNA Biology*, *14*(5), 637–650. https://doi.org/10.1080/15476286.2016.1201618

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, *9*(1), 171–181. https://doi.org/10.1038/nprot.2014.006

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, *10*(11), 1096–1100. https://doi.org/10.1038/nmeth.2639

Prostrate gland. (n.d.) *Farlex Partner Medical Dictionary*. (2012). Retrieved April 13 2019 from https://medical-dictionary.thefreedictionary.com/Prostrate+gland

Prostate Cancer Taskforce (2012). *Diagnosis and Management of Prostate Cancer in New Zealand Men: Recommendations from the Prostate Cancer Taskforce*. Wellington: Ministry of Health.

Qi, J., Pellecchia, M., & Ronai, Z. A. (2010). The Siah2-HIF-FoxA2 axis in prostate cancer – new markers and therapeutic opportunities. *Oncotarget*, *1*(5), 379–385. doi:10.18632/oncotarget.171

Quek, S. I., Ho, M. E., Loprieno, M. A., Ellis, W. J., Elliott, N., & Liu, A. Y. (2012). A Multiplex Assay to Measure RNA Transcripts of Prostate Cancer in Urine. *PLoS ONE*, *7*(9). https://doi.org/10.1371/journal.pone.0045656

Quek, S. I., Wong, O. M., Chen, A., Borges, G. T., Ellis, W. J., Salvanha, D. M., … Liu, A. Y. (2015). Processing of voided urine for prostate cancer RNA biomarker analysis. *Prostate*, *75*(16), 1886–1895. https://doi.org/10.1002/pros.23066

R Foundation for Statistical Computing. (2018). *R: a Language and Environment for Statistical Computing. http://www.R-project.org/*. Retrieved from http://www.r-project.org.

Raff, T., Van Der Giet, M., Endemann, D., Wiederholt, T., & Paul, M. (1997). Design and testing of β-actin primers for RT-PCR that do not co- amplify processed pseudogenes. *BioTechniques*, *23*(3), 456–460.

Ramsköld, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., Faridani, O. R., … Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, *30*(8), 777–782. https://doi.org/10.1038/nbt.2282

Raja, N., Russell, C. M., & George, A. K. (2018). Urinary markers aiding in the detection and risk stratification of prostate cancer. *Translational Andrology and Urology*, *7*(S4), S436–S442. https://doi.org/10.21037/tau.2018.07.01

Reynolds, A. P., Richards, G., De La Iglesia, B., & Rayward-Smith, V. J. (2006). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, *5*(4), 475–504. https://doi.org/10.1007/s10852-005-9022-1

Richie, J. P., Catalona, W. J., Ahmann, F. R., Hudson, M. A., Scardino, P. T., Flanigan, R. C., … Southwick, P. C. (1993). Effect of patient age on early detection of prostate cancer with serum prostate-specific antigen and digital rectal examination. *Urology*, *42*(4), 365–374. https://doi.org/10.1016/0090-4295(93)90359-I

Rifkin MD, Dahnert W, Kurtz AB (1990). State of the art: endorectal sonography of the prostate gland. AJR; 154:691-700

Rigau, M., Olivan, M., Garcia, M., Sequeiros, T., Montes, M., Colás, E., … Doll, A. (2013). The present and future of prostate cancer urine biomarkers. *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms140612620

Rigau, M., Ortega, I., Mir, M. C., Ballesteros, C., Garcia, M., Llauradõ, M., … Doll, A. (2011). A Three-Gene panel on urine increases PSA specificity in the detection of prostate cancer. *Prostate*, *71*(16), 1736–1745. https://doi.org/10.1002/pros.21390

Sakr, W. A., Haas, G. P., Cassin, B. F., Pontes, J. E., & Crissman, J. D. (1993). The frequency of carcinoma and intraepithelial neoplasia of the prostate in young male patients. *Journal of Urology*, *150*(2), 379–385. https://doi.org/10.1016/S0022-5347(17)35487-3

Salami, S. S., Schmidt, F., Laxman, B., Regan, M. M., Rickman, D. S., Scherr, D., … Sanda, M. G. (2013). Combining urinary detection of TMPRSS2: ERG and PCA3 with serum PSA to predict diagnosis of prostate cancer. *Urologic Oncology: Seminars and Original Investigations*, *31*(5), 566–571. https://doi.org/10.1016/j.urolonc.2011.04.001

Sartori, D. A., & Chan, D. W. (2014). Biomarkers in prostate cancer: what's new? *Current Opinion in Oncology*, *26*(3), 259–264. https://doi.org/10.1097/CCO.0000000000000065

Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K. D., Imai, T., & Ueda, H. R. (2013). Quartz-Seq: A highly reproducible and sensitive single-cell RNA sequencing method, reveals nongenetic gene-expression heterogeneity. *Genome Biology*, *14*(4). https://doi.org/10.1186/gb-2013-14-4-r31

Sharma, S. V., Lee, D. Y., Li, B., Quinlan, M. P., Takahashi, F., Maheswaran, S., … Settleman, J. (2010). A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. *Cell*, *141*(1), 69–80. https://doi.org/10.1016/j.cell.2010.02.027

Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, *69*(1), 7–34. https://doi.org/10.3322/caac.21551

Siegel, R., Naishadham, D., & Jemal, A. (2013). Cancer statistics, 2013 (US). *CA: A Cancer Journal for Clinicians*, *63*(1), 11–30. https://doi.org/10.3322/caac.21166

Smith, M., Lucia, M. S., Werahera, P. N., & La Rosa, F. G. (2010). Carcinoid tumor of the verumontanum (colliculus seminalis) of the prostatic urethra with a coexisting prostatic adenocarcinoma: A case report. *Journal of Medical Case Reports*, *4*. https://doi.org/10.1186/1752-1947-4-16

Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, *7*(3), 500–507. https://doi.org/10.1038/nprot.2011.457

Suslov, O. N., Kukekov, V. G., Laywell, E. D., Scheffler, B., & Steindler, D. a. (2000). RT-PCR amplification of mRNA from single brain neurospheres. *Journal of Neuroscience Methods*, *96*(1), 57–61. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10704671

Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, *13*, 599. Retrieved from https://doi.org/10.1038/nprot.2017.149

Tai, S., Sun, Y., Squires, J. M., Zhang, H., Oh, W. K., Liang, C. Z., & Huang, J. (2011). PC3 is a cell line characteristic of prostatic small cell carcinoma. *Prostate*, *71*(15), 1668–1679. https://doi.org/10.1002/pros.21383

Taplin, M. E., Bubley, G. J., Ko, Y. J., Small, E. J., Upton, M., Rajeshkumar, B., & Balk, S. P. (1999). Selection for androgen receptor mutations in prostate cancers treated with androgen antagonist. *Cancer Research*, *59*(11), 2511–2515.

The Health and Disability Commissioner. 2015. Delayed Diagnosis of Cancer in Primary Care: Complaints to the Health and Disability Commissioner: 2004 - 2013.

Udvardi, M. K., Czechowski, T., & Scheible, W.-R. (2008). Eleven Golden Rules of Quantitative RT-PCR. *THE PLANT CELL ONLINE*, *20*(7), 1736–1737. https://doi.org/10.1105/tpc.108.061143

Vlaeminck-Guillem, V., Ruffion, A., André, J., Devonec, M., & Paparel, P. (2010). Urinary Prostate Cancer 3 Test: Toward the Age of Reason? *Urology*. https://doi.org/10.1016/j.urology.2009.03.046

Wilhelm, J., Muyal, J. P., Best, J., Kwapiszewska, G., Stein, M. M., Seeger, W., … Fink, L. (2006). Systematic comparison of the T7-IVT and SMART-based RNA preamplification techniques for DNA microarray experiments. *Clinical Chemistry*, *52*(6), 1161–1167. https://doi.org/10.1373/clinchem.2005.062406

Woo, M.-O., Jeong, S.-C., Markkandan, K., Paek, N.-C., Choi, S.-B., & Seo, H. S. (2015). Isolation and Functional Studies of Genes. In *Current Technologies in Plant Molecular Breeding* (pp. 241–295). https://doi.org/10.1007/978-94-017-9996-6_8

Wu, C., Wyatt, A. W., Mcpherson, A., Lin, D., Mcconeghy, B. J., Mo, F., … Collins, C. C. (2012). Poly-gene fusion transcripts and chromothripsis in prostate cancer. *Genes Chromosomes and Cancer*, *51*(12), 1144–1153. https://doi.org/10.1002/gcc.21999

Yao, V., Berkman, C. E., Choi, J. K., O'Keefe, D. S., & Bacich, D. J. (2010). Expression of prostate-specific membrane antigen (PSMA), increases cell folate uptake and proliferation and suggests a novel role for PSMA in the uptake of the non-polyglutamated folate, folic acid. *Prostate*, *70*(3), 305–316. https://doi.org/10.1002/pros.21065

Yuan, T. C., Veeramani, S., Lin, F. F., Kondrikou, D., Zelivianski, S., Igawa, T., … Lin, M. F. (2006). Androgen deprivation induces human prostate epithelial neuroendocrine differentiation of androgen-sensitive LNCaP cells. *Endocrine-Related Cancer*, *13*(1), 151–167. https://doi.org/10.1677/erc.1.01043

Zajac, P., Islam, S., Hochgerner, H., Lönnerberg, P., & Linnarsson, S. (2013). Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. *PLoS ONE*, *8*(12). https://doi.org/10.1371/journal.pone.0085270

Zenklusen, D., Larson, D. R., & Singer, R. H. (2008). Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural and Molecular Biology*, *15*(12), 1263–1271. https://doi.org/10.1038/nsmb.1514

Zhu, Y., Horne, M. K., Manasseh, R., Aumann, T. D., Boon, W. C., & Petkovic-Duran, K. (2011). Increasing cDNA Yields from Single-cell Quantities of mRNA in Standard Laboratory Reverse Transcriptase Reactions using Acoustic Microstreaming. *Journal of Visualized Experiments*, (53). https://doi.org/10.3791/3144

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Górczak, K., Klamecka, K., Figlerowicz, M., & Siatkowski, I. (2015). The Impact of Normalization Methods on RNA-Seq Data Analysis. *BioMed Research International*, *2015*, 1–10. https://doi.org/10.1155/2015/621690